

THE THEORY OF NUMBERS

AN INTRODUCTION

ANTHONY A. GIOIA

THE THEORY OF NUMBERS

An Introduction

Anthony A. Gioia

DOVER PUBLICATIONS, INC.
Mineola, New York

Copyright

Copyright © 1970, 2001 by Anthony A. Gioia

All rights reserved under Pan American and International Copyright Conventions.

Published in Canada by General Publishing Company, Ltd., 30 Lesmill Road, Don Mills, Toronto, Ontario.

Bibliographical Note

This Dover edition, first published in 2001, is an unabridged reprint of *The Theory of Numbers: An Introduction*, originally published by Markham Publishing Co., Chicago, in 1970. The author has made a number of corrections for this new edition, as well as adding Solutions for Selected Exercises to the backmatter.

Library of Congress Cataloging-in-Publication Data

Gioia, Anthony A., 1934—

The theory of numbers : an introduction / Anthony A. Gioia.

p. cm. — (Markham mathematics series)

Originally published: Chicago : Markham Pub. Co., 1970.

Includes bibliographical references and index.

ISBN 0-486-41449-3 (pbk.)

1. Number theory. I. Title. II. Series.

QA241 .G56 2001

512'.7—dc21

00-059603

Manufactured in the United States of America
Dover Publications, Inc., 31 East 2nd Street, Mineola, N.Y. 11501

To Pat, Kathy, Mickey

PREFACE

THIS TEXT was developed for a first course in number theory at Western Michigan University, where students may take this course as early as their junior undergraduate year or as late as their first graduate year. With the diversity in the levels of the students' previous training in mind, I selected the topics and methods of presentation for this text. Every chapter is written under the assumption that the student's background consists only of calculus (including the brief introductions to analytic geometry and linear algebra). The instructor can achieve various depths in the course, in order to compensate for the differing backgrounds, by requiring individual students to complete exercises of a degree of difficulty proportionate to the student's proficiency.

Because of the assumption of only a post-calculus standard of maturity of the student, any of the chapters may be presented in any trimester regardless of variations in preparedness of the audience from one trimester to another. (To aid the instructor in the selection of chapters to be presented, a table of logical dependence of the sections is given at the end of this Preface.) Moreover, regardless of the extent of the student's prior training, his sophistication increases only slightly in one trimester; accordingly, the rate of presentation of material in this text remains almost constant throughout.

Arithmetic functions are introduced early (Chapter 2) and are used often in the development (Chapter 4; Chapter 8; Section 14 of Chapter 3; and Section 30 of Chapter 5). The emphasis is on the algebraic structure of the set of functions under Dirichlet convolution; although this has been common in the literature, it has received inadequate stress in published texts. The algebraic viewpoint has the obvious advantage of leading to the development of a calculus which eliminates the exigency of proofs replete with mysterious combinatorial and unpleasant computational techniques.

In Chapter 8 the Erdős-Selberg proof of the Prime Number Theorem is given. In part, my reason for choosing to give the elementary proof rather than a nonelementary proof of the Prime Number Theorem is that the elementary proof comes as an easy (though not short) application of methods already used in Chapter 4 to derive the standard results on orders of magnitude, namely, the Euler-McLaurin sum formula, summation by parts, and

changing the order of summations. And in part my reason is that the undergraduate rarely has the requisite background in function theory to follow a nonelementary proof.

In addition to the introductions to the theory of arithmetic functions and analytic number theory cited above, the text also gives introductions to algebraic and geometric number theory. The former is furnished by the study of the Gaussian integers (to enumerate the ways of representing a positive integer as a sum of two squares), and the Jacobian integers (to prove the insolubility of the Fermat cubic). Because of the emphasis on the similarity of methods used for the rational, Gaussian, and Jacobian integers, the student can appreciate that these are but three special cases of a general theory. The introduction to geometric number theory is supplied by the use of geometric methods in the proof of the Quadratic Reciprocity Law (the concept of a lattice point is introduced in Section 17); in the proofs of certain asymptotic formulas for summatory functions (in Sections 19, 21, and 22); and, of course, throughout Chapter 9.

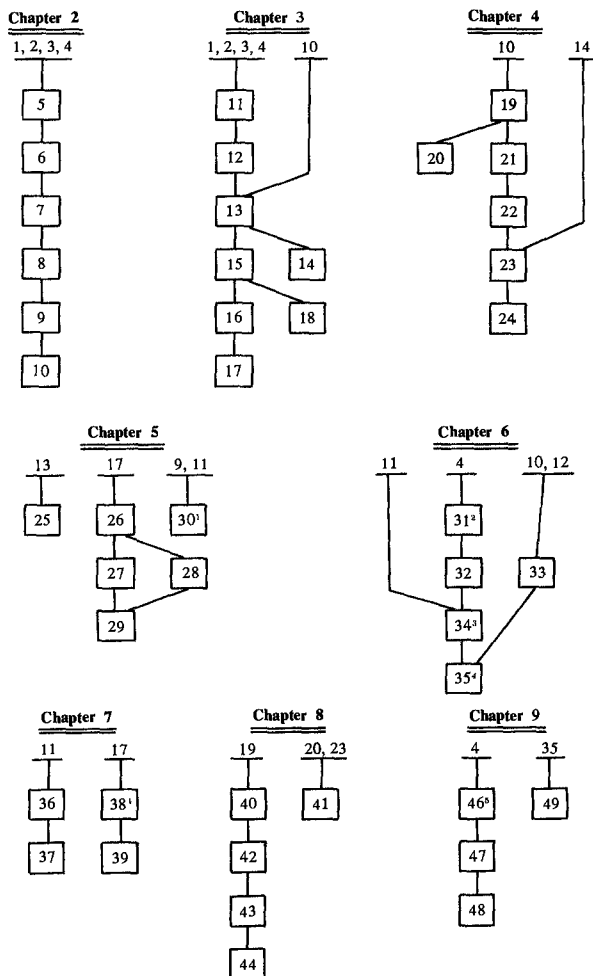
I wish to thank Professor John Petro and Dr. A. M. Vaidya for reading the manuscript, pointing out an embarrassing number of errors, and helping to supplement the original lists of exercises; Professor William J. LeVeque for his valuable suggestions about the final chapters; and Mrs. Judith Warriner for typing the manuscript.

A.A.G.

September, 1969

Kalamazoo, Michigan

LOGICAL DEPENDENCE OF SECTIONS



¹Need the statement of Theorem 27.1.

²Need Section 11 for Exercises.

³Need Section 17 for Exercises.

⁴Need Lemma 26.2, which can be proved after Section 10.

⁵Need Section 17 for an example.

CONTENTS

Preface

vii

CHAPTER 1 FUNDAMENTAL CONCEPTS

1.	Divisibility	1
2.	The gcd and the lcm	4
3.	The Euclidean Algorithm	7
4.	The Fundamental Theorem	9

CHAPTER 2 ARITHMETIC FUNCTIONS

5.	The Semigroup $\mathcal{A}$ of Arithmetic Functions	13
6.	The Group of Units in $\mathcal{A}$	16
7.	The Subgroup of Multiplicative Functions	18
8.	The Möbius Function and Inversion Formulas	20
9.	The Sigma Functions	23
10.	The Euler φ -Function	27

CHAPTER 3 CONGRUENCES AND RESIDUES

11.	Complete Residue Systems	31
12.	Linear Congruences	35
13.	Reduced Residue Systems	39
14.	Ramanujan's Trigonometric Sum	41
15.	Wilson's Theorem	45
16.	Primitive Roots	47
17.	Quadratic Residues	52
18.	Congruences with Composite Moduli	59

CHAPTER 4 SUMMATORY FUNCTIONS

19.	Introduction	63
20.	The Euler-McLaurin Sum Formula	70
21.	Order of Magnitude of $\tau(n)$	73
22.	Order of Magnitude of $\sigma(n)$	76
23.	Sums Involving the Möbius Function	78
24.	Squarefree Integers	82

CHAPTER 5 SUMS OF SQUARES

25.	Sums of Four Squares	84
26.	Sums of Two Squares	87
27.	Number of Representations	90
28.	The Gaussian Integers	90
29.	Proof of Theorem 27.1	95
30.	Restatement of Theorem 27.1	96

CHAPTER 6 CONTINUED FRACTIONS, FAREY SEQUENCES, THE PELL EQUATION

31.	Finite Continued Fractions	98
32.	Infinite Simple Continued Fractions	103
33.	Farey Sequences	107
34.	The Pell Equation	110
35.	Rational Approximations of Reals	114

CHAPTER 7 THE EQUATION $x^n + y^n = z^n$, $n \leq 4$

36.	Pythagorean Triples	120
37.	The Equation $x^4 + y^4 = z^4$	122
38.	Arithmetic in $\mathcal{K}(\sqrt{-3})$	123
39.	The Equation $x^3 + y^3 = z^3$	126

CHAPTER 8 THE PRIME NUMBER THEOREM

40.	Introductory Remarks	129
41.	Preliminary Results	130
42.	The Function $\psi(x)$	134
43.	A Fundamental Inequality	140
44.	The Behavior of $r(x)/x$	144
45.	The Prime Number Theorem and Related Results	153

CHAPTER 9 GEOMETRY OF NUMBERS

46.	Preliminaries	160
47.	Convex Symmetric Distance Functions	164
48.	The Theorems of Minkowski	170
49.	Applications to Farey Sequences and Continued Fractions	174

<i>Solutions for Selected Exercises</i>	181
<i>Bibliography</i>	203
<i>Index</i>	205

Chapter 1

FUNDAMENTAL CONCEPTS

1. Divisibility

Let $\mathcal{Z}$ denote the set of integers, that is,

$$\mathcal{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\},$$

and $\mathcal{Z}^+$ denote the set of positive integers, $\mathcal{Z}^+ = \{1, 2, 3, \dots\}$. The following definition of divisibility introduces a relation on $\mathcal{Z}$ which is basic to most of the material presented later.

Definition. Suppose $a, b \in \mathcal{Z}$. We say a divides b (or b is a multiple of a), and write $a|b$, if and only if there exists $c \in \mathcal{Z}$ such that $ac = b$. If a does not divide b , we write $a \nmid b$.

Notice that if $b \neq 0$, then $a|b$ implies that $a \neq 0$; equivalently, if $a = 0$ and $a|b$, then $b = 0$. However, if $b = 0$, then $a|b$ for every $a \in \mathcal{Z}$. Some further consequences of the definition are given in the theorems below.

Theorem 1.1.

- (a) For every $a \in \mathcal{Z}$, $a|a$.
- (b) If $a|b$ and $b|a$, then $a = \pm b$.
- (c) If $a|b$ and $b|c$, then $a|c$.
- (d) If $a|b$ and $a|c$, then $a|(bx + cy)$ for all $x, y \in \mathcal{Z}$.

Proof. The proofs of (a), (b), and (c) are left as exercises. To prove (d) we observe that $a|b$ and $a|c$ implies the existence of $d, e \in \mathcal{Z}$ such that $ad = b$ and $ae = c$. But then $a(dx + ey) = bx + cy$ for all $x, y \in \mathcal{Z}$, and since $dx + ey \in \mathcal{Z}$, we have $a|(bx + cy)$, as required.

Theorem 1.2. If $b \neq 0$ and $a|b$, then $|a| \leq |b|$.

Proof. For some $c \in \mathcal{Z}$, we have $ac = b$. Then $|a||c| = |b| \neq 0$. Thus, $|c| \neq 0$, so $|c| \geq 1$. Therefore, $|a| \leq |a||c| = |b|$.

Theorem 1.3. *If $a \in \mathcal{Z}$ and $b \in \mathcal{Z}^+$, then there exist unique $q, r \in \mathcal{Z}$ such that $a = bq + r$, $0 \leq r < b$.*

Proof. Consider the set

$$\mathcal{S} = \{a + bx : x \in \mathcal{Z}\}.$$

If $a \geq 0$, then $a + b \geq 0$ and $a + b \in \mathcal{S}$; if $a < 0$, then $a + b(-a) = a(1 - b) \geq 0$ and is an element of $\mathcal{S}$. In either case $\mathcal{S}$ contains non-negative integers, and therefore contains a smallest non-negative integer, say r . Suppose that r occurs in $\mathcal{S}$ with $x = -q$, so we have

$$0 \leq r = a - bq.$$

If $r \geq b$, then $a - bq \geq b$, from which it follows that

$$0 \leq a + b(-q - 1) < a - bq.$$

But this is impossible since $a - bq$ is the smallest non-negative integer of the form $a + bx$. Hence, $b > r$ and there are integers q and r as stated in the theorem.

To prove uniqueness, we assume there are integers q, q_1, r, r_1 , such that

$$a = bq + r, \quad 0 \leq r < b$$

and

$$a = bq_1 + r_1, \quad 0 \leq r_1 < b.$$

Now assume that $q < q_1$. Then

$$0 \leq r_1 = a - bq_1 \leq a - b(q + 1) = r - b < 0,$$

an obvious contradiction. Similarly, $q_1 \nless q$. Hence $q = q_1$, and

$$r_1 = a - bq_1 = a - bq = r.$$

Corollary 1.3. *(The division algorithm.) If $a \in \mathcal{Z}$ and $b \neq 0$, $b \in \mathcal{Z}$, then there exist unique $q, r \in \mathcal{Z}$ such that*

$$a = bq + r, \quad 0 \leq r < |b|.$$

Proof. If $b > 0$, the corollary is the same as Theorem 1.3. If $b < 0$, then a and $|b|$ satisfy the hypothesis of Theorem 1.3, so there exist $q', r \in \mathcal{Z}$ such that $a = |b|q' + r$, $0 \leq r < |b|$. But with $b < 0$, $|b| = -b$; take $q = -q'$ and the corollary is proved.

EXERCISES

1-1. Prove Theorem 1.1 (a), (b), and (c).

1-2. Show that Theorem 1.2 is not necessarily true if $b = 0$.

- 1-3. Suppose $\mathcal{S}$ is a set and $\mathcal{R}$ is a relation on $\mathcal{S}$. If $r, s \in \mathcal{S}$, we write $r\mathcal{R}s$ to mean that r is in the relation $\mathcal{R}$ to s . The relation $\mathcal{R}$ is called a *partial order* on $\mathcal{S}$ provided (a) $s\mathcal{R}s$ for all $s \in \mathcal{S}$, (b) if $r\mathcal{R}s$ and $s\mathcal{R}r$, then $r = s$, and (c) if $r\mathcal{R}s$ and $s\mathcal{R}t$, then $r\mathcal{R}t$. Prove that “ $\mid$ ” is a partial order on $\mathbb{Z}^+$.
- 1-4.* Prove by mathematical induction that if $a \mid b_1, a \mid b_2, \dots, a \mid b_n$, then $a \mid (b_1x_1 + \dots + b_nx_n)$ for every set $\{x_1, x_2, \dots, x_n\} \subset \mathbb{Z}$. Thus prove that if $a \mid b_i$ ($i = 1, \dots, n$) and if c can be written as a linear combination of the b_i (that is, there are integers $x_1, \dots, x_n$ such that $c = b_1x_1 + \dots + b_nx_n$), then $a \mid c$.
- 1-5.* Prove
- (a) If $a \mid b$, then $a \mid bc$ for every $c \in \mathbb{Z}$.
 - (b) If $a \mid b$, then $ac \mid bc$ for every $c \in \mathbb{Z}$.
 - (c) If $a \mid b$ and $c \mid d$, then $ac \mid bd$.
- 1-6.* If $a \neq 0$, show that the set $\{d : d \mid a\}$ is a finite set.
- 1-7.* Suppose $n \in \mathbb{Z}^+$ is fixed, and let

$$\mathcal{S} = \{d : d \in \mathbb{Z}^+, d \mid n\}.$$

Prove

- (a) $d \in \mathcal{S}$ if and only if $n/d \in \mathcal{S}$.
 - (b) If the elements of $\mathcal{S}$ are ordered by increasing magnitude, $1 = d_1 < d_2 < \dots < d_t = n$, then the corresponding elements n/d_i ($i = 1, \dots, t$) are ordered by decreasing magnitude.
- 1-8. Suppose $a, b \in \mathbb{Z}^+$ and $ab = c$. Prove that either $a \leq \sqrt{c}$ or $b \leq \sqrt{c}$.
- 1-9. An integer n is even if $2 \mid n$, odd if $2 \nmid n$. Prove that the sum and difference
- (a) of two even integers is even,
 - (b) of two odd integers is even,
 - (c) of an even and an odd integer is odd.
- 1-10. If n is an odd integer different from ± 1 , prove that n cannot divide both of two consecutive even integers, and that n cannot divide both of two consecutive odd integers.
- 1-11. Suppose $a, b, n \in \mathbb{Z}$ and $|a - b| < |n|$. Prove that n cannot divide both a and b , $a \neq b$.
- 1-12. Suppose $a \in \mathbb{Z}$. Prove
- (a) $a \mid n$ for every $n \in \mathbb{Z}$ if and only if $a = \pm 1$.
 - (b) $n \mid a$ for every $n \in \mathbb{Z}$ if and only if $a = 0$.
- 1-13. Let $a, b, c \in \mathbb{Z}$ and suppose $c \neq 0$. Show that $ac \mid bc$ implies $a \mid b$.

* An exercise marked with an asterisk contains a result which will be used in some later discussion.

2. The Greatest Common Divisor and the Least Common Multiple

Definition. Suppose $a, b \in \mathcal{Z}$, and not both of a and b are zero. The greatest common divisor (gcd) of a and b , denoted (a, b) , is an integer d satisfying

- (a) $d \in \mathcal{Z}^+$,
- (b) $d|a$ and $d|b$,
- (c) if $e|a$ and $e|b$, then $e|d$.

For example, $(4, 12) = 4$, $(2, 3) = 1$, $(6, 8) = 2$. It is not obvious that a gcd of two integers, not both zero, always exists, but this will be proved in Theorem 2.1. However, it is easy to see that if a gcd exists, then it is unique, because if $d = (a, b)$ and $e = (a, b)$, then $e|a$ and $e|b$ by (b); hence, $e|d$ by (c). Similarly $d|e$. But from (a), both d and e are positive, so $d = e$.

Theorem 2.1. If a and b are not both zero, then the gcd (a, b) exists, is the least positive member of the set

$$\mathcal{S} = \{au + bv : u, v \in \mathcal{Z}\},$$

and (a, b) divides every element of $\mathcal{S}$. In particular, there exist $x, y \in \mathcal{Z}$ such that $(a, b) = ax + by$.

Proof. Clearly the integers $\pm a = a(\pm 1) + b(0)$ and $\pm b = a(0) + b(\pm 1)$ are all in $\mathcal{S}$ and at least one of $\pm a$ or $\pm b$ is positive, so $\mathcal{S}$ contains positive elements. Therefore, $\mathcal{S}$ contains a least positive integer, say d . Suppose $d = au_0 + bv_0$. We will prove that $d = (a, b)$.

Let $c = au_1 + bv_1$ be any element of $\mathcal{S}$. There exist q, r such that $c = dq + r$, $0 \leq r < d$. Now the integer r is in $\mathcal{S}$, since

$$\begin{aligned} r &= c - dq = (au_1 + bv_1) - (au_0 + bv_0)q \\ &= a(u_1 - u_0q) + b(v_1 - v_0q). \end{aligned}$$

But $r < d$ implies $r = 0$ since d is the least positive element of $\mathcal{S}$. Therefore, $c = dq$; we have proved that $d|c$ for every $c \in \mathcal{S}$.

Since $a, b \in \mathcal{S}$, we have $d|a$ and $d|b$. Thus, d has the properties (a) and (b) of the gcd, and we now show that d also has property (c). Suppose $e|a$ and $e|b$. By Theorem 1.1(d), e divides any linear combination of a and b . But d is such a combination, so $e|d$, and it follows that $d = (a, b)$. The proof is complete.

We can say a little more about the set $\mathcal{S}$ of Theorem 2.1. We showed above that if $c \in \mathcal{S}$, then $d|c$; the converse is also true, namely, if c is any multiple of d , then $c \in \mathcal{S}$, because if $d|c$, then $dn = c$ for some $n \in \mathcal{Z}$. We have $d = ax + by$ from the theorem, so $c = dn = a(xn) + b(yn) \in \mathcal{S}$. This proves the following corollary.

Corollary 2.1a. Suppose $a, b, c \in \mathcal{Z}$, a and b are not both zero, and $d = (a, b)$. Then there exist integers u and v such that $c = au + bv$ if and only if $d|c$.

If $(a, b) = 1$, then a and b are called *relatively prime*. If a and b are relatively prime, then clearly there are integers u and v such that $au + bv = 1$. Conversely, if $au + bv = 1$ for some $u, v \in \mathcal{Z}$, then by Corollary 2.1a, $(a, b)|1$, and since $0 < (a, b)$ we must have $(a, b) = 1$. Therefore we have the following.

Corollary 2.1b. The relative primality of a and b is necessary and sufficient for the existence of $u, v \in \mathcal{Z}$ such that $au + bv = 1$.

Some typical facts relating the concepts introduced above are given in the next theorems.

Theorem 2.2. If $(a, b) = d$, then $(a/d, b/d) = 1$.

Proof. Let $d = (a, b)$. Then there exist x and y such that $d = ax + by$. Therefore

$$1 = \frac{a}{d}x + \frac{b}{d}y.$$

But a/d and $b/d \in \mathcal{Z}$ so, by Corollary 2.1b, $(a/d, b/d) = 1$.

Theorem 2.3. If $a|bc$ and $(a, b) = 1$, then $a|c$.

Proof. From $(a, b) = 1$ we have $ax + by = 1$ for some $x, y \in \mathcal{Z}$. Then $acx + bcy = c$. But $a|ac$ and $a|bc$, so by Theorem 1.1(d), $a|c$.

Definition. Suppose $a, b \in \mathcal{Z}$, $ab \neq 0$. The least common multiple (lcm) of a and b , denoted $\langle a, b \rangle$, is an integer m defined by

- (a) $m \in \mathcal{Z}^+$,
- (b) $a|m$ and $b|m$,
- (c) if $a|n$ and $b|n$, then $m|n$.

For example, $\langle 4, 12 \rangle = 12$, $\langle 2, 3 \rangle = 6$, $\langle 6, 8 \rangle = 24$. It is easy to see that $\langle a, b \rangle$ always exists for any pair of non-zero integers a and b , and in fact the lcm is the least positive member of the set

$$\mathcal{S} = \{n : n \in \mathcal{Z}^+, a|n, b|n\}.$$

The essential part of the argument should be familiar by now. Since $|ab| \in \mathcal{S}$ there is a least positive member $m \in \mathcal{S}$. Clearly m has properties (a) and (b). Suppose $n \in \mathcal{S}$ and $a|n, b|n$. We find $q, r \in \mathcal{Z}$ such that $n = mq + r$, $0 \leq r < m$. Now $a|n, a|m$, so $a|r$. Similarly, $b|r$, so either $r \in \mathcal{S}$ or $r = 0$. By the way m was chosen, $r = 0$. Therefore $m|n$, and m has property (c) so that $m = \langle a, b \rangle$.

There is a fundamental relationship between the gcd and the lcm of two non-zero integers which is given in Theorem 2.4.

Theorem 2.4. Suppose $ab \neq 0$. Then $(a,b)\langle a,b \rangle = |ab|$.

Proof. Suppose $(a,b) = d$ and $\langle a,b \rangle = m$. Our method of proof will be to show that $ab|dm$ and $dm|ab$. We first observe that $d|ab$ (because $d|a$, so $d|ab$) and that $m|ab$ (since $a|ab$ and $b|ab$, the definition of the lcm gives $m|ab$), or what is the same thing, that $ab/d \in \mathcal{L}$ and $ab/m \in \mathcal{L}$.

Now $d|b$, so $a d|ab$, or $d|(ab/d)$. Similarly, $b|(ab/d)$, so $m|(ab/d)$, or $dm|ab$. But $b|m$, so $ab|am$, or $(ab/m)|a$. Similarly, $(ab/m)|b$, so $(ab/m)|d$, or $ab|dm$. Therefore, $dm = \pm ab = |ab|$.

EXERCISES

2-1.* Prove

- (a) $(ac, bc) = |c|(a, b)$;
- (b) if $a|c$, $b|c$, and $(a, b) = 1$, then $ab|c$;
- (c) $(a, b) = |a|$ if and only if $a|b$;
- (d) if $(a, b) = 1$ and $(a, c) = 1$, then $(a, bc) = 1$.

2-2. Prove

- (a) $\langle ac, bc \rangle = |c|\langle a, b \rangle$;
- (b) $\langle a, b \rangle = |a|$ if and only if $b|a$.

2-3. Prove

- (a) if $(a, b) = (a, c)$ and $\langle a, b \rangle = \langle a, c \rangle$, then $b = \pm c$;
- (b) $\langle a, (b, c) \rangle = (\langle a, b \rangle, \langle a, c \rangle)$;
- (c) $(a, \langle b, c \rangle) = \langle (a, b), (a, c) \rangle$.

2-4.* Suppose not both of a, b are zero. Let

$$\mathcal{D}_1 = \{d: d|a\} \quad \text{and} \quad \mathcal{D}_2 = \{d: d|b\}.$$

Show that $\mathcal{D}_1 \cap \mathcal{D}_2$ is a non-empty finite set. Prove that (a, b) is the largest element in $\mathcal{D}_1 \cap \mathcal{D}_2$.

2-5. Suppose $f(x)$ is a polynomial of degree ≥ 1 with coefficients in $\mathcal{L}$. If the equation $f(x) = 0$ has a rational root a/b , $b \neq 0$, $(a, b) = 1$, prove that a divides the constant term of $f(x)$ and b divides the leading coefficient of $f(x)$.

2-6. Prove this partial converse of Theorem 2.2: If $0 < d$, $d|a$, $d|b$, and $(a/d, b/d) = 1$, then $(a, b) = d$.

2-7. Suppose $ab \neq 0$. Let

$$\mathcal{E}_1 = \{e: a|e\} \quad \text{and} \quad \mathcal{E}_2 = \{e: b|e\}.$$

Prove that $\langle a, b \rangle$ is the smallest positive element in $\mathcal{E}_1 \cap \mathcal{E}_2$.

2-8.* The gcd of more than two integers is defined as follows: If $a_1, \dots, a_n$ ($n \geq 2$) are not all zero, then their gcd $(a_1, \dots, a_n) = d$ is the unique positive integer such that $d|a_i$ ($i = 1, \dots, n$), and if $e|a_i$ ($i = 1, \dots, n$), then $e|d$. Prove by induction that for $n > 2$, $(a_1, \dots, a_n) = ((a_1, \dots, a_{n-1}), a_n)$.

- 2-9.* The lcm of more than two integers all different from zero is the unique positive integer $m = \langle a_1, \dots, a_n \rangle$ such that $a_i | m$ ($i = 1, \dots, n$), and if $a_i | M$ ($i = 1, \dots, n$), then $m | M$. Prove that for $n > 2$, $\langle a_1, \dots, a_n \rangle = \langle \langle a_1, \dots, a_{n-1} \rangle, a_n \rangle$.
- 2-10. If $(a, b) = 1$, we know there are integers u and v such that $au + bv = 1$. Prove $(u, v) = (b, u) = (a, v) = 1$.
- 2-11.* If $b \neq 0$ and $a = bq + r$, $0 \leq r < |b|$, prove $(a, b) = (b, r)$.
- 2-12. An *integral module* is a set $\mathcal{M}$ such that (a) $\emptyset \neq \mathcal{M} \subset \mathcal{Z}$, and (b) if $a \in \mathcal{M}$, $b \in \mathcal{M}$, then $a + b \in \mathcal{M}$ and $a - b \in \mathcal{M}$. Prove that if $\mathcal{M}$ is any integral module, there exists $m \in \mathcal{Z}$ (m depends on $\mathcal{M}$) such that $\mathcal{M} = \{0, \pm m, \pm 2m, \pm 3m, \dots\}$.
- 2-13. Suppose $\mathcal{S}$ is a set such that (a) $\emptyset \neq \mathcal{S} \subset \mathcal{Z}$, and (b) if $a \in \mathcal{S}$, $b \in \mathcal{S}$, then $a - b \in \mathcal{S}$. Prove that $\mathcal{S}$ is an integral module (see Exercise 2-12).
- 2-14. If $n \in \mathcal{Z}^+$ is fixed, consider the n (not necessarily distinct) integers d_i defined by $d_i = (i, n)$, $i = 1, \dots, n$. Prove that the set $\{d_1, \dots, d_n\}$ is just the set $\mathcal{S}$ of Exercise 1-7.
- 2-15. Give an example to show that the solution x, y of the equation $(a, b) = ax + by$ is not unique.
- 2-16. Find necessary and sufficient conditions on a and b so that $(a, b) = \langle a, b \rangle$.
- 2-17. Prove that for every $n \in \mathcal{Z}$, $(3, n^2 + 1) = 1$, $(5, n^2 + 2) = 1$, and $(7, n^2 + 2) = 1$.
- 2-18. Suppose $\{a_1, a_2, \dots\}$ is an infinite sequence of integers satisfying $a_{n+2} = a_{n+1} + a_n$ ($n \geq 1$). If $(a_1, a_2) = 1$, prove $(a_n, a_{n+1}) = 1$ for all $n \in \mathcal{Z}^+$. If $(a_1, a_2) = d$, prove $d | (a_n, a_{n+1})$ for all $n \in \mathcal{Z}^+$.
- 2-19. If $a, b, n \in \mathcal{Z}^+$ and $n > 1$, find the gcd and the lcm of $n^a - 1$ and $n^b - 1$.

3. The Euclidean Algorithm

We have proved the existence and uniqueness of the gcd of two integers not both zero. Moreover, implicit in Exercise 2-4 is a method for finding the gcd, though that method is inefficient and cumbersome for very large a, b . We now consider a more efficient method for finding the gcd of two integers.

Let us begin with an example. To find $(245, 1022)$ we repeatedly use the division algorithm to find

- (1) $1022 = 245(4) + 42, \quad 0 \leq 42 < 245$
- (2) $245 = 42(5) + 35, \quad 0 \leq 35 < 42$
- (3) $42 = 35(1) + 7, \quad 0 \leq 7 < 35$
- (4) $35 = 7(5) + 0, \quad 0 \leq 0 < 7.$

Now we can say that $7 = (245, 1022)$, because from (4), $7|35$; therefore $7|7$ and $7|35$, so 7 divides any linear combination of 7 and 35. In particular, 42 is such a combination, from (3), so $7|42$. Now $7|35$ and $7|42$, so $7|245$ since (2) shows 245 is a linear combination of 35 and 42. Finally, we see from (1) that $7|1022$ since $7|42$ and $7|245$. Hence, 7 is a common divisor of 245 and of 1022.

To show that 7 is the greatest of the common divisors, suppose $e|245$ and $e|1022$. First we rewrite (1), (2), and (3) in the form

$$(5) \quad 42 = 1022 + 245(-4)$$

$$(6) \quad 35 = 245 + 42(-5)$$

$$(7) \quad 7 = 42 + 35(-1).$$

From (5) we see that $e|42$; then from (6) we see that $e|35$. Consequently, (7) tells us that $e|7$, so that 7 is the gcd. This entire argument may be given more elegantly by using Exercise 2-11 and Exercise 2-1(c) to observe that $(1022, 245) = (245, 42) = (42, 35) = (35, 7) = 7$.

This process for finding the gcd is known as the *Euclidean algorithm*. We can also use this algorithm to find integers x and y such that $7 = 1022x + 245y$, as follows. Substitute from (6) into (7) to get

$$\begin{aligned} 7 &= 42 + 35(-1) \\ &= 42 + (245 + 42(-5))(-1) \\ &= 245(-1) + 42(6); \end{aligned}$$

now substitute from (5) to get

$$\begin{aligned} 7 &= 245(-1) + (1022 + 245(-4))(6) \\ &= 1022(6) + 245(-25). \end{aligned}$$

The above algorithm may be applied to any pair of integers a, b not both zero. Since $(a, 0) = |a|$ for every $a \neq 0$, and since $(a, b) = (|a|, |b|)$, we may assume that both a and b are positive. Then we use the division algorithm repeatedly to find q_i, r_i ($i = 1, \dots, k+1$) such that

$$\begin{aligned} a &= bq_1 + r_1, & 0 < r_1 < b \\ b &= r_1q_2 + r_2, & 0 < r_2 < r_1 \\ r_1 &= r_2q_3 + r_3, & 0 < r_3 < r_2 \\ &\dots & \\ r_{k-2} &= r_{k-1}q_k + r_k, & 0 < r_k < r_{k-1} \\ r_{k-1} &= r_kq_{k+1} + r_{k+1}, & 0 = r_{k+1} < r_k. \end{aligned}$$

Then $(a,b) = r_k$, the last non-zero remainder. It is clear that the process must terminate with a remainder zero after a finite number of steps, because $r_1 > r_2 > \cdots \geq 0$ is a *decreasing* sequence of non-negative integers and r_{k+1} must be zero for some k .

Also, as in the example, the sequence of equations can be used to find x,y such that $r_k = ax + by$. The lcm of two integers can be found by using Theorem 2.4. In the exercises below the student will show that the gcd of more than two integers can be found by using Exercise 2-8 and the Euclidean algorithm; the lcm can be found by using Exercise 2-9 and Theorem 2.4.

EXERCISES

- 3-1. Find $(20,35)$, $(112,96)$, $(27,45)$.
- 3-2. Use the Euclidean algorithm to find integers x,y such that $(20,35) = 20x + 35y$; $(112,96) = 112x + 96y$; $(27,45) = 27x + 45y$.
- 3-3. Find $\langle 20,35 \rangle$, $\langle 112,96 \rangle$, $\langle 27,45 \rangle$.
- 3-4. Use Exercise 2-8 and the Euclidean algorithm to find
 - (a) $(60,30,42,8)$,
 - (b) $(42,60,8,30)$,
 - (c) $(2250,30,540,900)$.
- 3-5. Use Exercise 2-9 and Theorem 2.4 to find
 - (a) $\langle 60,30,42,8 \rangle$,
 - (b) $\langle 42,60,8,30 \rangle$,
 - (c) $\langle 2250,30,540,900 \rangle$.
- 3-6. Suppose k is a non-zero integer and m and n are distinct positive integers. Prove that

$$(2k)^{2^n} + 1 \quad \text{and} \quad (2k)^{2^m} + 1$$

are relatively prime.

4. The Fundamental Theorem

Definition. A positive integer p is called a prime if and only if p has exactly two (distinct) positive divisors. An integer $n > 1$ which is not prime is called composite.

We note that if $n \in \mathcal{L}$ and p is prime, then (n,p) is either 1 or p , and the latter occurs if and only if $p|n$.

Theorem 4.1. Suppose p is prime and $a_1, \dots, a_k \in \mathcal{L}$. If $p|a_1 \cdots a_k$, then $p|a_i$ for some i , $1 \leq i \leq k$.

Proof. We prove this theorem by induction on k . For $k = 1$ there is nothing to prove. For $k = 2$, assume $p|a_1 a_2$; if $p|a_1$ we are through, so suppose $p \nmid a_1$. Then $(p, a_1) = 1$, and by Theorem 2.3 we have $p|a_2$.

Now assume the theorem has been proved for $k = 1, \dots, n-1$ ($2 \leq n$) and suppose $p|a_1 \cdots a_n$. We have $p|(a_1 \cdots a_{n-1})a_n$; therefore, either $p|a_n$ or $p|a_1 \cdots a_{n-1}$. If $p|a_n$ we are through. If not, then $p|a_1 \cdots a_{n-1}$, and by the inductive assumption, $p|a_i$ for some i , $1 \leq i \leq n-1$. In any case, then, if $p|a_1 \cdots a_n$, then $p|a_i$ for some i , $1 \leq i \leq n$, and the proof is complete.

Theorem 4.2. (*Fundamental theorem of number theory.*) *If $n > 1$, then n can be represented as a product of a finite number of primes, and this representation is unique except for the order of the prime factors.*

Proof. (Existence.) For $n = 2$, we obviously have n expressed as a product of primes. Suppose for all n , $2 \leq n \leq N-1$, it has been shown that n is a product of a finite number of primes. If N is itself a prime, we are through. If N is not a prime, then it has a divisor d , $1 < d < N$, say $de = N$. But then $1 < e < N$, and the inductive hypothesis guarantees that both d and e have a prime factor decomposition, and hence so does N .

(Uniqueness.) We again proceed by induction. For $n = 2$, the factorization is unique. Suppose uniqueness has been proved for all n , $2 \leq n < N$. If N is prime, its factorization is unique, so suppose N is composite. We will assume that N has at least two representations as a product of primes and reach a contradiction.

Suppose $N = p_1 \cdots p_r$ and $N = q_1 \cdots q_s$, where p_i ($i = 1, \dots, r$) and q_j ($j = 1, \dots, s$) are primes. Clearly r and s are both larger than 1 since N is not prime. From

$$p_1 \cdots p_r = N = q_1 \cdots q_s$$

we have $p_1|q_1 \cdots q_s$, so $p_1|q_k$ for some k , $1 \leq k \leq s$ by Theorem 4.1. Since $1 < p_1$, and q_k has exactly two positive divisors, we know $p_1 = q_k$.

But then $N/p_1 < N$ and $1 < p_2 \cdots p_r = N/p_1$, so that $N/p_1 = N/q_k$ has a unique representation as a product of primes. Therefore, $r = s$, and for each $i = 2, \dots, r$, it follows that $p_i = q_j$ for some j , $j \neq k$, $1 \leq j \leq s$. Since $N/p_1 = p_2 \cdots p_r$ uniquely, then $N = p_1 p_2 \cdots p_r$ is the unique factorization of N . This completes the induction.

From the factorization of $n > 1$ into primes, we obtain a representation

$$(8) \quad n = \prod_{i=1}^r p_i^{\alpha_i}$$

where the p_i are distinct primes and the $\alpha_i \in \mathcal{L}^+$. Often it is convenient to allow $\alpha_i \geq 0$. The factorization (8) is called the *canonical* form of n . For example, $15 = 3^1 \cdot 5^1$ and $25 = 5^2$, though sometimes it might be desirable to write $25 = 3^0 \cdot 5^2$. Thus we could say that the canonical forms for 15 and 25 formally involve the same primes.

EXERCISES

4-1. Suppose

$$1 < n = \prod_{i=1}^r p_i^{\alpha_i}, \quad \alpha_i > 0.$$

Prove that for $d > 0$, $d|n$ if and only if

$$d = \prod_{i=1}^r p_i^{\beta_i}$$

with $0 \leq \beta_i \leq \alpha_i$ for $i = 1, \dots, r$.

4-2.* Suppose

$$a = \prod_{i=1}^r p_i^{\alpha_i}, \quad \alpha_i \geq 0, \quad \text{and} \quad b = \prod_{i=1}^r p_i^{\beta_i}, \quad \beta_i \geq 0.$$

Prove that

$$(a, b) = \prod_{i=1}^r p_i^{\gamma_i}$$

where $\gamma_i = \min(\alpha_i, \beta_i)$, $i = 1, \dots, r$. Hence, use Theorem 2.4 to conclude that

$$\langle a, b \rangle = \prod_{i=1}^r p_i^{\delta_i}$$

where $\delta_i = \max(\alpha_i, \beta_i)$.

4-3.* If $(m, n) = 1$ and $0 < d|mn$, prove that d has a unique factorization $d = ab$ such that $a > 0$, $b > 0$, $a|m$, $b|n$, $(a, b) = 1$ and $(m/a, n/b) = 1$. Prove that $a = (d, m)$ and $b = (d, n)$.

4-4. Suppose $k > 2$ and $a_1, \dots, a_k \in \mathcal{Z}^+$. If

$$a_j = \prod_{i=1}^r p_i^{\alpha_{ji}}, \quad \alpha_{ji} \geq 0 \quad (j = 1, \dots, k),$$

prove that

$$(a_1, \dots, a_k) = \prod_{i=1}^r p_i^{\beta_i}$$

where $\beta_i = \min(\alpha_{1i}, \dots, \alpha_{ki})$ for $i = 1, \dots, r$.

4-5. With the same notation as in Exercise 4-4, prove that

$$\langle a_1, \dots, a_k \rangle = \prod_{i=1}^r p_i^{\delta_i}$$

where $\delta_i = \max(\alpha_{1i}, \dots, \alpha_{ki})$ for $i = 1, \dots, r$.

- 4-6. Notice that $(2,4,6) = 2$ and that $\langle 2,4,6 \rangle = 12$, so that $24 = (2,4,6) \times \langle 2,4,6 \rangle \neq 2 \cdot 4 \cdot 6$. Thus, Theorem 2.4 cannot be extended without some restrictions. Find necessary and sufficient conditions so that for $k > 2$,

$$(a_1, \dots, a_k) \langle a_1, \dots, a_k \rangle = |a_1 \cdots a_k|.$$

- 4-7.* Suppose $p_1, \dots, p_n$ are primes and let $m = p_1 \cdots p_n + 1$. Prove that m has a prime divisor p such that $p \neq p_1, \dots, p \neq p_n$. Hence prove that there are infinitely many primes.
- 4-8. Suppose $\mathcal{S}$ is a set containing $n + 1$ integers t such that $1 \leq t \leq 2n$ for every $t \in \mathcal{S}$. Prove there exist integers $a, b \in \mathcal{S}$, $a \neq b$, such that $a|b$. ["Elementary Problem E1765," *Amer. Math. Monthly*, Vol. 72 (1965), p. 183.]
- 4-9. Use the results of Exercise 4-2 to prove
- $\langle a, (b, c) \rangle = (\langle a, b \rangle, \langle a, c \rangle)$,
 - $(a, \langle b, c \rangle) = \langle (a, b), (a, c) \rangle$.
- 4-10. Let $\mathcal{S} = \{3n - 2 : n \in \mathbb{Z}^+\}$. Define a *prime in $\mathcal{S}$* to be any element of $\mathcal{S}$ which has exactly two distinct divisors in $\mathcal{S}$. For example, 4 and 7 are primes in $\mathcal{S}$.
- Prove that the product of two elements in $\mathcal{S}$ is in $\mathcal{S}$.
 - Given any $t \in \mathcal{S}$, prove that either $t = 1$, t is a prime in $\mathcal{S}$, or t can be written as a product of primes in $\mathcal{S}$.
 - Show that 10 and 25 are primes in $\mathcal{S}$.
 - Notice that $4 \cdot 25 = 10 \cdot 10$, so that factorization into primes in $\mathcal{S}$ is not unique. Find another element of $\mathcal{S}$ which has at least two different factorizations into primes in $\mathcal{S}$.

- 4-11. Consider the table

4	7	10	13	16	19	...
7	12	17	22	27	32	...
10	17	24	31	38	45	...
13	22	31	40	49	58	...
.....						

in which the rule of formation is evident. Show that a positive integer k occurs in the table if and only if $2k + 1$ is not a prime.

- 4-12. Find a prime which is simultaneously of each of the forms $a^2 + b^2$, $a^2 + 2b^2, \dots, a^2 + 10b^2$ for some $a, b \in \mathbb{Z}$.

Chapter 2

ARITHMETIC FUNCTIONS

5. The Semigroup $\mathcal{A}$ of Arithmetic Functions.

Definition. An arithmetic function is a function whose domain is $\mathcal{Z}^+$ and whose range is a subset of the set of complex numbers.

For example, $f(n) = n$ for all $n \in \mathcal{Z}^+$, and $g(n) = e^{in}$ where $i^2 = -1$ are arithmetic functions. Let $\mathcal{A}$ denote the set of all arithmetic functions. If f and $g \in \mathcal{A}$, the product (or Dirichlet convolution product) of f and g is the function denoted by $f \cdot g$ and defined at each $n \in \mathcal{Z}^+$ by

$$f \cdot g(n) = \sum_{d|n} f(d) g\left(\frac{n}{d}\right)$$

where the summation extends over all positive divisors d of n .

Examples.

$$f \cdot g(1) = f(1) g(1)$$

$$f \cdot g(2) = f(1) g(2) + f(2) g(1)$$

$$f \cdot g(6) = f(1) g(6) + f(2) g(3) + f(3) g(2) + f(6) g(1)$$

Notice that this product is in fact a binary operation on $\mathcal{A}$ since $f, g \in \mathcal{A}$ implies $f \cdot g \in \mathcal{A}$, which follows from the fact that sums and products of complex numbers are complex numbers. It will be convenient to have the following equivalent form of the definition of $f \cdot g$.

Lemma 5.1. If $f, g \in \mathcal{A}$, then for every $n \in \mathcal{Z}^+$,

$$f \cdot g(n) = \sum_{ab=n} f(a) g(b)$$

where the summation extends over all ordered pairs a, b of positive integers whose product is n .

Proof. By Exercise 1-7, $d|n$ if and only if $(n/d)|n$. Therefore, as d ranges over all positive divisors of n , the ordered pairs $d, n/d$ range over all ordered pairs of positive integers whose product is n .

Theorem 5.1. *The product of arithmetic functions is associative; that is, if $f, g, h \in \mathcal{A}$, then $(f \cdot g) \cdot h(n) = f \cdot (g \cdot h)(n)$ for every $n \in \mathcal{Z}^+$.*

Proof. For any fixed $n \in \mathcal{Z}^+$,

$$\begin{aligned} (f \cdot g) \cdot h(n) &= \sum_{dc=n} f \cdot g(d) h(c) \\ &= \sum_{dc=n} \left\{ \sum_{ab=d} f(a) g(b) \right\} h(c) \\ &= \sum_{abc=n} f(a) g(b) h(c) \\ &= \sum_{ae=n} f(a) \left\{ \sum_{bc=e} g(b) h(c) \right\} \\ &= \sum_{ae=n} f(a) g \cdot h(e) \\ &= f \cdot (g \cdot h)(n). \end{aligned}$$

An algebraic system $(\mathcal{S}, \times)$ consisting of a set $\mathcal{S}$ together with an operation “ $\times$ ” on $\mathcal{S}$ is called a *semigroup* provided that “ $\times$ ” is associative. In this terminology we have proved that $(\mathcal{A}, \cdot)$ is a semigroup. This semigroup of arithmetic functions has other interesting properties, as we now show.

Theorem 5.2. *The product of arithmetic functions is a commutative operation; that is, if $f, g \in \mathcal{A}$, then $f \cdot g(n) = g \cdot f(n)$ for all $n \in \mathcal{Z}^+$.*

Proof.

$$\begin{aligned} f \cdot g(n) &= \sum_{ab=n} f(a) g(b) \\ &= \sum_{ba=n} g(b) f(a) \\ &= g \cdot f(n). \end{aligned}$$

We now consider this question: Does there exist a function ε in $\mathcal{A}$ such that $f \cdot \varepsilon = \varepsilon \cdot f = f$ for every $f \in \mathcal{A}$? The reader is probably aware of the fact that such an element in a semigroup is called the *identity* of the semigroup. Thus, our question is: Does there exist an identity ε in $(\mathcal{A}, \cdot)$? Evidently, since the product of functions is commutative, it will suffice to investigate the existence of a function ε such that $f = \varepsilon \cdot f$ for all $f \in \mathcal{A}$, or equivalently, such that for every $f \in \mathcal{A}$,

$$(1) \quad f(n) = \varepsilon \cdot f(n) \quad \text{for all } n \in \mathcal{Z}^+.$$

We shall assume there is such an identity ε and find the necessary properties which this function must have. Writing out (1) for $n = 1$, we have $f(1) = \varepsilon(1)f(1)$. Since this last equation must hold for all $f \in \mathcal{A}$, it must hold in particular if $f(1) \neq 0$, so we must have $\varepsilon(1) = 1$. Writing out (1) for $n = 2$ gives

$$f(2) = \varepsilon(1)f(2) + \varepsilon(2)f(1).$$

Since $\varepsilon(1) = 1$, the above equation reduces to $0 = \varepsilon(2)f(1)$, which will hold for arbitrary f only if $\varepsilon(2) = 0$.

Now we continue by induction. Assume that $\varepsilon(n) = 0$ for $n = 2, \dots, k-1$ ($k \geq 3$) and consider (1) for $n = k$.

$$f(k) = \sum_{d|k} \varepsilon(d) f\left(\frac{k}{d}\right)$$

$$f(k) = \varepsilon(1)f(k) + \sum_{\substack{d|k \\ 1 < d < k}} \varepsilon(d) f\left(\frac{k}{d}\right) + \varepsilon(k)f(1)$$

Any terms appearing in the summation will vanish by the inductive assumption; also $\varepsilon(1) = 1$, so the above equation is simply $0 = \varepsilon(k)f(1)$. Thus, $\varepsilon(k) = 0$ is necessary, and by induction we have: If ε is the identity in $(\mathcal{A}, \cdot)$, then $\varepsilon(1) = 1$ and $\varepsilon(n) = 0$ if $n > 1$. The converse of this statement answers the question asked earlier. The proof of the converse is left as an exercise.

A convenient description of the function ε can be given in terms of the *largest integer* function (sometimes called the “square bracket” function). If x is a real number, then $[x]$ is defined by

$$[x] = \max \{n : n \in \mathcal{Z}, n \leq x\}.$$

Thus, $[x]$ is the largest integer which does not exceed x ; for example, $[2] = 2$, $[\sqrt{3}] = 1$, $[-7/2] = -4$. Now notice that $\varepsilon(n) = [1/n]$.

The results of this section are summarized in Theorem 5.3.

Theorem 5.3. *The system $(\mathcal{A}, \cdot)$ is a commutative semigroup with identity ε , where ε is the arithmetic function satisfying $\varepsilon(n) = [1/n]$.*

EXERCISES

- 5-1.* Prove that ε is the identity in $(\mathcal{A}, \cdot)$.
- 5-2. Suppose $f(n) = 1$ for all $n \in \mathcal{Z}^+$, and suppose p is a prime. For $k \in \mathcal{Z}^+$, find $f \cdot f(p^k)$.
- 5-3. Suppose $f(2) = -1$ and $f(n) = 0$ if $n \neq 2$, $g(3) = 1$ and $g(n) = 0$ if $n \neq 3$. Find $f \cdot g$.

- 5-4. Suppose $f(n) = 1$ for every $n \in \mathcal{Z}^+$, $g(1) = 1$, $g(p) = -1$, and $g(p^k) = 0$ for $k > 1$, where p is a fixed prime. Find $f \cdot f \cdot g(p^k)$ for all $k \geq 0$.
- 5-5.* Define $\theta(n) = 0$ for every $n \in \mathcal{Z}^+$. Show that $\theta \cdot f = \theta$ for every $f \in \mathcal{A}$, and if $g \cdot f = g$ for every $f \in \mathcal{A}$, then $g = \theta$.
- 5-6. An element f in a semigroup is called an *idempotent* if it has the property that $f \cdot f = f$. Show that ε and θ are the only idempotents in $\mathcal{A}$.

6. The Group of Units in $\mathcal{A}$

Definition. Suppose $f \in \mathcal{A}$. If there is a function $f' \in \mathcal{A}$ such that $f \cdot f' = \varepsilon$, then f' is called the inverse of f .

Evidently not every arithmetic function has an inverse. For example, there can be no function θ' (see Exercise 5-5 for the definition of θ) such that $\theta \cdot \theta'(n) = \varepsilon(n)$ for every n , because at $n = 1$ this would require that

$$0 = \theta(1) \theta'(1) = \theta \cdot \theta'(1) = \varepsilon(1) = 1,$$

which is impossible. The example shows essentially the only way that a function can fail to have an inverse. The situation is described in the next theorem.

Theorem 6.1. A necessary and sufficient condition that the inverse of f exist is that $f(1) \neq 0$.

Proof. Suppose f' exists. Since $f \cdot f'(1) = \varepsilon(1)$, we have $f(1) \neq 0$.

Conversely, suppose $f(1) \neq 0$ and consider the function g defined inductively by

$$\begin{aligned} g(1) &= \frac{1}{f(1)} \\ g(n) &= \frac{-1}{f(1)} \sum_{\substack{cd=n \\ 1 < c}} f(c) g(d), \quad n > 1. \end{aligned}$$

We will show that $f \cdot g = \varepsilon$, which will prove that $g = f'$ (see Exercise 6-1). Obviously $f \cdot g(1) = 1 = \varepsilon(1)$ and

$$\begin{aligned} f \cdot g(2) &= f(1) g(2) + f(2) g(1) \\ &= f(1) \left\{ \frac{-1}{f(1)} f(2) g(1) \right\} + f(2) g(1) \\ &= 0 = \varepsilon(2). \end{aligned}$$

Suppose $f \cdot g(k) = 0$ for $k = 2, \dots, n-1$ ($n \geq 3$). Then

$$\begin{aligned}
 f \cdot g(n) &= f(n)g(1) + \sum_{\substack{ab=n \\ 1 < b}} f(a)g(b) \\
 &= f(n)g(1) + \sum_{\substack{ab=n \\ 1 < b}} f(a) \left\{ \frac{-1}{f(1)} \sum_{\substack{cd=b \\ 1 < c}} f(c)g(d) \right\} \\
 &= f(n)g(1) - \frac{1}{f(1)} \sum_{\substack{ab=n \\ 1 < b}} f(a) \left\{ \sum_{cd=b} f(c)g(d) - f(1)g(b) \right\} \\
 &= f(n)g(1) - \frac{1}{f(1)} \sum_{\substack{ab=n \\ 1 < b}} f(a) \sum_{cd=b} f(c)g(d) + \sum_{\substack{ab=n \\ 1 < b}} f(a)g(b) \\
 &= f(n)g(1) - \frac{1}{f(1)} \sum_{\substack{ab=n \\ 1 < b}} f(a)f \cdot g(b) + \sum_{\substack{ab=n \\ 1 < b}} f(a)g(b).
 \end{aligned}$$

By the inductive assumption, the only non-zero term in the first summation occurs for $b = n$, so the above simplifies to

$$f \cdot g(n) = f(n)g(1) - \frac{1}{f(1)} f(1) f \cdot g(n) + \sum_{\substack{ab=n \\ 1 < b}} f(a)g(b) = 0.$$

This completes the induction and shows that for all n , $f \cdot g(n) = \varepsilon(n)$.

Definition. The set $\mathcal{U} \subset \mathcal{A}$ is defined by

$$\mathcal{U} = \{f \in \mathcal{A} : f(1) \neq 0\}.$$

The last theorem shows that $\mathcal{U}$ is just the set of functions which have inverses. This set $\mathcal{U}$ is called the *set of units* in $\mathcal{A}$. It is known that the set of units in a semigroup is a group called the “group of units”; though this result holds in more general situations, we state it only for $\mathcal{U}$. The proof is left as an exercise.

Theorem 6.2. The set $\mathcal{U}$ of units in $\mathcal{A}$ is a commutative group. That is,

- (a) if $f, g \in \mathcal{U}$, then $f \cdot g \in \mathcal{U}$;
- (b) the product of functions in $\mathcal{U}$ is associative and commutative;
- (c) the identity $\varepsilon \in \mathcal{U}$;
- (d) if $f \in \mathcal{U}$, then f' exists and $f' \in \mathcal{U}$.

EXERCISES

- 6-1. Suppose $f \in \mathcal{A}$. If there exists a pair of functions f' and $g \in \mathcal{A}$ such that $f \cdot f' = \varepsilon$ and $f \cdot g = \varepsilon$, prove that $f' = g$.
- 6-2. Prove Theorem 6.2.

- 6-3. If $f(n) = 1$ for all n , find $f'(n)$ for $1 \leq n \leq 10$.
 6-4. If $f(n) = 1$ for all n and p is any prime, find $f'(p)$ and $f'(p^2)$.
 6-5. Prove that $\varepsilon' = \varepsilon$.
 6-6. If f' exists, prove that $(f')'$, the inverse of f' , exists and that $(f')' = f$.

7. The Subgroup of Multiplicative Functions

Definition. Let $\mathcal{M}' = \{f \in \mathcal{A} : \text{if } (m, n) = 1, \text{ then } f(mn) = f(m)f(n)\}$ and $\mathcal{M} = \{f \in \mathcal{M}' : f(n) \neq 0 \text{ for some } n \in \mathcal{Z}^+\}$. A function f is called multiplicative if and only if $f \in \mathcal{M}$.

It is clear that $\mathcal{M}' = \mathcal{M} \cup \{\theta\}$ where $\theta \in \mathcal{A}$ is the function defined in Exercise 5-5. We also note that $f \in \mathcal{M}$ if and only if $f \in \mathcal{M}'$ and $f(1) = 1$, because if $f \in \mathcal{M}'$ and $f(1) = 1 \neq 0$, clearly $f \in \mathcal{M}$; conversely, suppose $f \in \mathcal{M}$ and let $n_0 \in \mathcal{Z}^+$ be such that $f(n_0) \neq 0$. Then $f \in \mathcal{M}'$, and since $(1, n_0) = 1$, $f(1 \cdot n_0) = f(1)f(n_0)$ implies $f(1) = 1$.

From the above remarks it is obvious that $\mathcal{M} \subset \mathcal{U}$. The set $\mathcal{M}$ of multiplicative functions is in fact a subgroup of the group of units, as we now show.

Theorem 7.1. If $f, g \in \mathcal{M}$, then $f \cdot g \in \mathcal{M}$.

Proof. (In this proof we use the result of Exercise 4-3.) If $f, g \in \mathcal{M}$, then $f(1) = 1 = g(1)$ and it follows that $f \cdot g(1) = 1$. If $(m, n) = 1$, then

$$\begin{aligned} f \cdot g(mn) &= \sum_{d|mn} f(d) g\left(\frac{mn}{d}\right) \\ &= \sum_{\substack{a|m \\ b|n \\ ab=d}} f(ab) g\left(\frac{m}{a} \frac{n}{b}\right) \\ &= \sum_{\substack{a|m \\ b|n}} f(a) f(b) g\left(\frac{m}{a}\right) g\left(\frac{n}{b}\right) \\ &= \left\{ \sum_{a|m} f(a) g\left(\frac{m}{a}\right) \right\} \left\{ \sum_{b|n} f(b) g\left(\frac{n}{b}\right) \right\} \\ &= f \cdot g(m) f \cdot g(n). \end{aligned}$$

Thus, $f \cdot g \in \mathcal{M}$.

The above theorem may be applied to obtain an interesting corollary. We first introduce a special class of multiplicative functions.

Definition. Let s be any complex number. The functions $\iota_s \in \mathcal{A}$ are defined by

$$\iota_s(n) = n^s, \quad n \in \mathcal{Z}^+.$$

We will refer to these functions as the *iota functions*.

We have left as an exercise the demonstration that every iota function is multiplicative—in particular, the function $\iota_0 \in \mathcal{M}$ [$\iota_0(n) = 1$ for all n]; it follows from Theorem 7.1 that if $f \in \mathcal{M}$, then $f \cdot \iota_0 \in \mathcal{M}$. This is equivalent to the following important result.

Corollary 7.1. *If f is a multiplicative function and if*

$$g(n) = \sum_{d|n} f(d),$$

then g is also multiplicative.

Continuing now with the demonstration that $\mathcal{M}$ is a group, we establish the following.

Theorem 7.2. *If $f \in \mathcal{M}$, then f' exists and $f' \in \mathcal{M}$.*

Proof. If $f \in \mathcal{M} \subset \mathcal{U}$, clearly f' exists. Also $f'(1) = 1/f(1) = 1$. Suppose we have shown that $f'(st) = f'(s)f'(t)$ whenever $(s,t) = 1$ for all s, t such that $1 \leq st \leq k-1$. Let $k = mn$ be any factorization of k such that $(m,n) = 1$. Now consider

$$\begin{aligned} 0 = \varepsilon(k) &= \sum_{d|k} f(d) f'\left(\frac{k}{d}\right) \\ &= \sum_{\substack{a|m \\ b|n}} f(ab) f'\left(\frac{m}{a} \frac{n}{b}\right) \\ &= \sum_{\substack{a|m \\ b|n \\ ab > 1}} f(ab) f'\left(\frac{m}{a} \frac{n}{b}\right) + f(1) f'(mn) \\ &= \sum_{\substack{a|m \\ b|n \\ ab > 1}} f(a) f(b) f'\left(\frac{m}{a}\right) f'\left(\frac{n}{b}\right) + f'(mn) \\ &= \sum_{\substack{a|m \\ b|n}} f(a) f(b) f'\left(\frac{m}{a}\right) f'\left(\frac{n}{b}\right) - f'(m) f'(n) + f'(mn) \\ &= \left\{ \sum_{a|m} f(a) f'\left(\frac{m}{a}\right) \right\} \left\{ \sum_{b|n} f(b) f'\left(\frac{n}{b}\right) \right\} - f'(m) f'(n) + f'(mn) \end{aligned}$$

$$(2) \quad 0 = \varepsilon(m) \varepsilon(n) - f'(m) f'(n) + f'(mn).$$

Since $1 < k$, at least one of m, n is larger than 1, so $\varepsilon(m) \varepsilon(n)$ vanishes. From (2) we have $f'(m) f'(n) = f'(mn)$, and by induction, $f' \in \mathcal{M}$.

We have already shown that the product of functions is associative and commutative in $\mathcal{A}$; therefore, the product also has these properties in $\mathcal{M}$. It is easy to see that the identity $\varepsilon \in \mathcal{M}$. These observations, together with our last two theorems, prove Theorem 7.3.

Theorem 7.3. *The set $\mathcal{M}$ of multiplicative functions is a commutative group.*

EXERCISES

7-1. Show that ε is multiplicative.

7-2.* If $n \in \mathcal{Z}^+$ and s is a complex number, say $s = \sigma + it$ with σ, t real, then $n^s = n^\sigma \{\cos(t \log n) + i \sin(t \log n)\}$. If $m, n \in \mathcal{Z}^+$ and s is complex, prove that $m^s n^s = (mn)^s$. Hence prove that for every complex $s, t_s \in \mathcal{M}$.

7-3. Give an example of a function f such that $f(1) = 1$ but f is not multiplicative.

8. The Möbius Function and Inversion Formulas

The next theorem is of help in studying particular functions in $\mathcal{M}$.

Theorem 8.1. *Suppose $f \in \mathcal{M}$ and $n > 1$ has the canonical form*

$$(3) \quad n = \prod_{i=1}^k p_i^{\beta_i}, \quad \beta_i > 0.$$

Then

$$f(n) = \prod_{i=1}^k f(p_i^{\beta_i}).$$

Proof. Let f and n be as in the hypothesis. We use induction on k . For $k = 1$ there is nothing to prove, so assume the theorem is true for $1 \leq k < K$. Then

$$\begin{aligned} f\left(\prod_{i=1}^K p_i^{\beta_i}\right) &= f\left(\left\{\prod_{i=1}^{K-1} p_i^{\beta_i}\right\} p_K^{\beta_K}\right) \\ &= f\left(\prod_{i=1}^{K-1} p_i^{\beta_i}\right) f(p_K^{\beta_K}) \\ &= \left\{\prod_{i=1}^{K-1} f(p_i^{\beta_i})\right\} f(p_K^{\beta_K}) \\ &= \prod_{i=1}^K f(p_i^{\beta_i}). \end{aligned}$$

This theorem tells us that a multiplicative function is completely determined by its values at p^β for every prime p and for every $\beta \in \mathcal{Z}^+$. We will show an application of Theorem 8.1 by finding the inverse of $\iota_0(n) = n^0 = 1$ for every n . Since $\iota_0 \in \mathcal{M}$, we know by Theorem 7.2 that ι'_0 exists and $\iota'_0 \in \mathcal{M}$, so we need only find $\iota'_0(p^\beta)$ for primes p and for $\beta = 1, 2, \dots$. Therefore, let p be any fixed prime; we proceed by induction on β . For $\beta = 1$ we have

$$\begin{aligned} 0 &= \varepsilon(p) = \iota'_0 \cdot \iota_0(p) \\ &= \iota'_0(1) \iota_0(p) + \iota'_0(p) \iota_0(1) \\ &= 1 + \iota'_0(p), \end{aligned}$$

from which we see that $\iota'_0(p) = -1$.

For $\beta = 2$, $\varepsilon(p^2) = \iota'_0 \cdot \iota_0(p^2)$ yields

$$0 = \sum_{j=0}^2 \iota'_0(p^j) \iota_0(p^{2-j}) = 1 - 1 + \iota'_0(p^2),$$

so $\iota'_0(p^2) = 0$. Now assume that $\iota'_0(p^\beta) = 0$ for $2 \leq \beta < B$; then

$$\begin{aligned} 0 &= \varepsilon(p^B) = \sum_{j=0}^B \iota'_0(p^j) \\ &= \iota'_0(1) + \iota'_0(p) + \sum_{j=2}^{B-1} \iota'_0(p^j) + \iota'_0(p^B) \\ &= \iota'_0(p^B). \end{aligned}$$

We have shown that $\iota'_0(p) = -1$ and $\iota'_0(p^\beta) = 0$ if $\beta \geq 2$ for every prime p .

Traditionally this function ι'_0 is called the *Möbius function*, and is denoted by μ . Adopting this notation, we apply Theorem 8.1 to conclude that if $n > 1$ is factored as in (3), then

$$\mu(n) = \prod_{i=1}^k \mu(p_i^{\beta_i}),$$

which is 0 if any of the β_i is at least 2, or $(-1)^k$ if every $\beta_i = 1$.

The discussion is summarized in Theorem 8.2.

Theorem 8.2. *The Möbius function μ , defined by $\iota_0 \cdot \mu = \mu \cdot \iota_0 = \varepsilon$, is a multiplicative function with values given by*

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1; \\ (-1)^k & \text{if } n \text{ is the product of } k \text{ distinct primes;} \\ 0 & \text{if } n \text{ is divisible by the square of some prime.} \end{cases}$$

Equivalently, the Möbius function satisfies the equations $\mu(1) = 1$, and for $n > 1$,

$$\sum_{d|n} \mu(d) = 0.$$

Corollary 8.2a. (Möbius inversion formula.) If

$$(4) \quad g(n) = \sum_{d|n} f(d),$$

then

$$(5) \quad f(n) = \sum_{d|n} g(d) \mu\left(\frac{n}{d}\right).$$

Conversely, (5) implies (4).

Proof. The equation (4) is simply $g = f \cdot \iota_0$, and (5) is $f = g \cdot \mu$. Evidently $g = f \cdot \iota_0$ if and only if $g \cdot \mu = f \cdot \iota_0 \cdot \mu = f \cdot \varepsilon = f$.

Corollary 8.2b. Suppose

$$g(n) = \sum_{d|n} f(d).$$

If g is multiplicative, then so is f .

Proof. This corollary is the converse of Corollary 7.1. If $g = f \cdot \iota_0$, then $g \cdot \mu = f$. Now g and $\mu \in \mathcal{M}$, so their product $f \in \mathcal{M}$.

The Möbius inversion formula suggests the derivation of a wide class of inversion formulas, which we make in our next theorem. The proof of the theorem is trivial. Notice that the Möbius inversion formula is a special case of the theorem.

Theorem 8.3. Suppose $h \in \mathcal{U}$ and $f, g \in \mathcal{A}$. Then $f \cdot h = g$ if and only if $f = g \cdot h'$.

EXERCISES

8-1. Prove Theorem 8.3.

8-2. An integer n is called *squarefree* if for every $t > 1$, $t^2 \nmid n$.

(a) Prove that the integer 1 is the only positive integer which is both squarefree and a square.

(b) Show that n is squarefree if and only if $\mu(n) \neq 0$.

(c) A function $f \in \mathcal{A}$ is called the *characteristic function* of a set $\mathcal{S} \subset \mathcal{X}^+$ if $f(n) = 1$ when $n \in \mathcal{S}$ and $f(n) = 0$ when $n \notin \mathcal{S}$. Prove that the characteristic function of the set of positive squarefree integers is μ^2 ($\mu^2(n) = (\mu(n))^2$).

8-3. Use the methods of this section to find ι'_1 , the inverse of ι_1 .

- 8-4.* For any complex number s , show that the inverse ι'_s of the corresponding iota function ι_s is the multiplicative function such that $\iota'_s(p^\beta) = -p^s$ if $\beta = 1$, and $= 0$ if $\beta > 1$, for every prime p .
- 8-5. For which of the following sets is the characteristic function of the set a multiplicative function?
- $\{1, p, p^2\}$, prime p
 - $\{1, p^2\}$, prime p
 - $\{2, 3, 6\}$
 - $\{1, 2, 3, 6\}$
 - $\{1, 4, 6, 24\}$
 - $\{1, 3, 4, 12\}$
- 8-6. Let f be the characteristic function of a set $\mathcal{S} \subset \mathcal{Z}^+$, $\mathcal{S} \neq \emptyset$. Show that $f \in \mathcal{M}$ if and only if both of the following conditions hold:
- $1 \in \mathcal{S}$;
 - if $(a, b) = 1$, $a \in \mathcal{S}$ and $b \in \mathcal{S}$ if and only if $ab \in \mathcal{S}$.
- 8-7. Define the functions $F_j(n) = (j, n)$ for $j = 0, 1, 2$.
- Which of the functions F_0, F_1, F_2 are in $\mathcal{U}$?
 - Which of the functions F_j are in $\mathcal{M}$?
 - Are any of the functions F_j iota functions?
 - Is any one of F_0, F_1, F_2 the characteristic function of some set $\mathcal{S} \subset \mathcal{Z}^+$?
 - Find $F_2 \cdot \mu(3^\beta)$ for every $\beta \in \mathcal{Z}^+$.
- 8-8. If

$$n = \prod_{i=1}^k p_i^{\beta_i},$$

define $\rho(n) = \beta_1 + \cdots + \beta_k$ and $v(n) = k$, $v(1) = \rho(1) = 0$. Let $\lambda(n) = (-1)^{\rho(n)}$ and $\eta(n) = 2^{v(n)}$. Prove

$$\sum_{d|n} \lambda(d) \mu(d) = \eta(n), \quad \text{and} \quad \sum_{d|n} \lambda(d) \mu\left(\frac{n}{d}\right) = \eta(n) \lambda(n).$$

(Hint: η and $\lambda \in \mathcal{M}$; use Theorem 8.1.)

- 8-9. Use the definitions given in Exercise 8-8 and prove that

$$\sum_{ab=n} (-1)^{\rho(a)} 2^{v(b)} = 1;$$

hence prove that $\lambda \cdot \eta = \iota_0$.

9. The Sigma Functions

Definition. If s is any complex number, the function $\sigma_s \in \mathcal{A}$ is defined by $\sigma_s = \iota_s \cdot \iota_0$. The functions σ_s are called sigma functions.

Note. If $s = 1$ we will write σ instead of σ_1 ; if $s = 0$ we will write τ in place of σ_0 . These exceptional notations for the sigma functions corresponding to $s = 0$ and $s = 1$ are introduced so that our notation will agree with established conventions. However, it is usually convenient to discuss the class of sigma functions without treating σ and τ separately.

It is immediate from Theorem 7.1 that every sigma function is multiplicative. Also

$$\begin{aligned}\sigma_s(n) &= \iota_s \cdot \iota_0(n) \\ &= \sum_{d|n} \iota_s(d) \iota_0\left(\frac{n}{d}\right) \\ &= \sum_{d|n} d^s,\end{aligned}$$

from which it is clear that $\sigma_s(n)$ denotes the sum of the s^{th} powers of the positive divisors of n . In particular, $\sigma(n)$ is the sum of the divisors of n , and $\tau(n)$ is the number of divisors of n . Furthermore, if p is prime,

$$\sigma_s(p^\beta) = \sum_{j=0}^{\beta} (p^j)^s = \begin{cases} \beta + 1 & \text{if } s = 0, \\ \frac{p^{s(\beta+1)} - 1}{p^s - 1} & \text{if } s \neq 0. \end{cases}$$

Now Theorem 8.1 may be used to show that if

$$1 < n = \prod_{j=1}^k p_j^{\beta_j},$$

then

$$\begin{aligned}\tau(n) &= \prod_{j=1}^k (\beta_j + 1), \\ \sigma(n) &= \prod_{j=1}^k \frac{p_j^{\beta_j+1} - 1}{p_j - 1}, \\ \sigma_s(n) &= \prod_{j=1}^k \frac{p_j^{s(\beta_j+1)} - 1}{p_j^s - 1}, \quad s \neq 0.\end{aligned}$$

It is easy to derive various identities involving these arithmetic functions. For example,

(a) by commutativity we have

$$\sigma_s(n) = \iota_s \cdot \iota_0(n) = \sum_{d|n} \left(\frac{n}{d}\right)^s = n^s \sum_{d|n} \frac{1}{d^s};$$

(b) by Möbius inversion, $\sigma_s \cdot \mu = \iota_s$ or

$$n^s = \sum_{d|n} \sigma_s(d) \mu\left(\frac{n}{d}\right);$$

(c) since the inverse of the product of functions is the product of the inverses,

$$\sigma'_s = (\iota_s \cdot \iota_0)' = \iota'_s \cdot \iota'_0 = \iota'_s \cdot \mu,$$

and ι'_s was found in Exercise 8-4.

Some classical problems, still not completely solved, have been stated in terms of the σ -function. We will discuss one of these, the problem of *perfect numbers*. Prior to discussing perfect numbers we define a *Mersenne prime* to be any prime q of the form $q = a^b - 1$ where $a, b \in \mathbb{Z}^+$ and $b > 1$. For example, $3 = 2^2 - 1$ and $7 = 2^3 - 1$ are Mersenne primes. By using the identity

$$x^{n+1} - 1 = (x - 1)(x^n + x^{n-1} + \cdots + 1),$$

it is easy to see that if $q = a^b - 1$ is a Mersenne prime, then $a = 2$ and b is prime, because $(a - 1)|(a^b - 1)$, and since q is prime, $a - 1 = 1$. Similarly, if b is composite, then b has a divisor d , $1 < d < b$, and $(2^d - 1)|q$, contradicting the primality of q . Thus, a Mersenne prime is a prime of the form $2^p - 1$, where p is prime.

A positive integer n is called *perfect* if $\sigma(n) = 2n$. It is not known whether there are any odd perfect integers, but the following theorem gives a complete characterization of all even perfect numbers. The sufficiency was known to Euclid; the necessity was first proved by Euler.

Theorem 9.1. *A necessary and sufficient condition that n be an even perfect number is that $n = 2^{p-1}(2^p - 1)$, where $2^p - 1$ is a Mersenne prime.*

Proof. Suppose $n = 2^{p-1}(2^p - 1)$, with p and $2^p - 1$ prime. Then $\sigma(n) = \sigma(2^{p-1})\sigma(2^p - 1) = 2n$.

Conversely, suppose n is even and perfect. Say $n = 2^k m$, $k \geq 1$ and m odd. We have $\sigma(n) = (2^{k+1} - 1)\sigma(m) = 2n = 2^{k+1}m$. Then $(2^{k+1} - 1)|m$, say $m = (2^{k+1} - 1)M$. Substituting above, we get $\sigma(m) = 2^{k+1}M$. Now m and M are distinct divisors of m , so $\sigma(m) \geq m + M$. But this gives

$$2^{k+1}M = \sigma(m) \geq m + M = 2^{k+1}M,$$

so that m and M are the only divisors of m . Hence $M = 1$ and $m = 2^{k+1} - 1$ is prime; as we saw earlier, $2^{k+1} - 1$ is prime only if $k + 1 = p$ is prime. Thus n is of the form described in the theorem.

EXERCISES

9-1. Prove that $\tau(n)$ is odd if and only if n is a square.

9-2. Prove that $\sigma(n)$ is odd if and only if n is a square or two times a square.

9-3. Prove that for all $n \in \mathcal{Z}^+$

$$(a) \sum_{d|n} \tau(d) \mu\left(\frac{n}{d}\right) = 1,$$

$$(b) \sum_{d|n} \sigma(d) \mu\left(\frac{n}{d}\right) = n.$$

9-4. Suppose r and s are complex numbers. Prove that $\sigma_s \cdot \sigma_r = \tau \cdot \iota_s \cdot \iota_r$. Write this identity using the summation notation involved in the definition of the product of functions.

9-5.* Prove that the square of the sum of the first n positive integers is equal to the sum of the cubes of the first n positive integers. If $f \in \mathcal{A}$, define f^k by $f^k(n) = (f(n))^k$. Prove $f \in \mathcal{M}$ implies $f^k \in \mathcal{M}$ for all real k . Combine these results to prove that $(\iota_0 \cdot \iota_0 \cdot \iota_0)^2 = (\tau \cdot \iota_0)^2 = \tau^3 \cdot \iota_0 = (\iota_0 \cdot \iota_0)^3 \cdot \iota_0$.

9-6. If $m, n \in \mathcal{Z}^+$, then m and n are called *amicable* if $\sigma(m) = \sigma(n) = m + n$. (For example, 220 and 284 are amicable.) Prove that if m is even, n odd, and m and n are amicable, then m is either a square or twice a square, and n is a square.

9-7. Suppose

$$f(n) = \prod_{d|n} g(d).$$

Take logarithms and apply Möbius inversion to prove that this is equivalent to

$$g(n) = \prod_{d|n} f^{\mu(d)}\left(\frac{n}{d}\right).$$

9-8. Find all solutions of the equation $\tau(n) = 10$, $n \in \mathcal{Z}^+$.

9-9. Prove that $3|\sigma(3n+2)$ and $4|\sigma(4n+3)$ for all $n \geq 0$, $n \in \mathcal{Z}$.

9-10. Prove that $12|\sigma(12n-1)$ for all $n \in \mathcal{Z}^+$.

9-11. If p is prime, prove $\sigma(p^a) \mid \sigma(p^b)$ if and only if $(a+1) \mid (b+1)$.

9-12. Prove that

$$\sigma(n) = \sum_{m=1}^n \int_0^m \cos\left(\frac{2\pi n[x+1]}{m}\right) dx.$$

9-13. Prove that $\tau(x) = n$ has infinitely many solutions x for every $n > 1$.

9-14. Show that

$$\prod_{d|n} d = n^{\tau(n)/2}, \quad n \in \mathcal{Z}^+.$$

9-15. Prove that for $n > 1$,

$$\prod_{d|n} d = n^2$$

if and only if $n = p^3$ or $n = pq$, where p, q are primes.

10. The Euler φ -Function

Definition. The function φ , called the Euler φ -function, is defined by $\varphi = \iota_1 \cdot \mu$.

Since ι_1 and $\mu \in \mathcal{M}$, it follows that $\varphi \in \mathcal{M}$. We proceed to find $\varphi(p^\beta)$ for prime p .

$$\begin{aligned} \varphi(p^\beta) &= \sum_{d|p^\beta} \iota_1(d) \mu\left(\frac{p^\beta}{d}\right) \\ &= \sum_{j=0}^{\beta} \iota_1(p^{\beta-j}) \mu(p^j) \\ &= \iota_1(p^\beta) \mu(1) + \iota_1(p^{\beta-1}) \mu(p) \\ &= p^\beta - p^{\beta-1} = p^\beta \left(1 - \frac{1}{p}\right) \end{aligned}$$

Then if

$$\begin{aligned} 1 < n &= \prod_{j=1}^r p_j^{\beta_j}, \\ \varphi(n) &= \prod_{j=1}^r p_j^{\beta_j} \left(1 - \frac{1}{p_j}\right) = n \prod_{j=1}^r \left(1 - \frac{1}{p_j}\right). \end{aligned}$$

If $\mathcal{S}$ is a finite set, we write $v_{\mathcal{S}}$ to denote the number of elements in $\mathcal{S}$. Consider the function γ defined at each $n \in \mathcal{Z}^+$ by

$$\gamma(n) = v\{t: 1 \leq t \leq n, \quad (t, n) = 1\}.$$

We will prove that

$$\sum_{d|n} \gamma\left(\frac{n}{d}\right) = n.$$

Let $n \in \mathcal{Z}^+$ be arbitrary and fixed. We define

$$\mathcal{S} = \{1, 2, \dots, n\}.$$

Now for each d , $1 \leq d \leq n$, we define

$$\mathcal{S}_d = \{t: 1 \leq t \leq n, \quad (t, n) = d\}.$$

We first observe that the collection $\{\mathcal{S}_d: 1 \leq d \leq n\}$ is a *partition* of $\mathcal{S}$, that is, a collection of disjoint subsets of $\mathcal{S}$, and every element of $\mathcal{S}$ is in some $\mathcal{S}_d$. If $d \neq e$, $1 \leq d \leq n$ and $1 \leq e \leq n$, suppose $\mathcal{S}_e \cap \mathcal{S}_d \neq \emptyset$. Then there is an integer t , $1 \leq t \leq n$, such that $t \in \mathcal{S}_e \cap \mathcal{S}_d$. But this implies $e = (t, n) = d$, which is impossible. Therefore, if $d \neq e$, then $\mathcal{S}_d \cap \mathcal{S}_e = \emptyset$. Also, if $t \in \mathcal{S}$, let $d = (t, n)$. Since $d|n$ and $1 \leq d \leq n$, we have $t \in \mathcal{S}_d$, so every element of $\mathcal{S}$ is in some one of the subsets.

Since the collection of subsets $\mathcal{S}_d$ is a partition, the number of elements in $\mathcal{S}$ is the sum of the numbers of elements in $\mathcal{S}_d$, $d = 1, \dots, n$. Notice, however, that if $d \nmid n$, then $\mathcal{S}_d = \emptyset$, because if $t \in \mathcal{S}_d$, then $(t, n) = d$ and $d|n$. Therefore, $v\mathcal{S}_d = 0$ if $d \nmid n$, and

$$n = v\mathcal{S} = \sum_{d=1}^n v\mathcal{S}_d = \sum_{d|n} v\mathcal{S}_d.$$

To complete the proof, we notice that for every d such that $d|n$,

$$\begin{aligned} v\mathcal{S}_d &= v\{t: 1 \leq t \leq n, (t, n) = d\} \\ &= v\left\{\frac{t}{d}: 1 \leq \frac{t}{d} \leq \frac{n}{d}, \left(\frac{t}{d}, \frac{n}{d}\right) = 1\right\} = \gamma\left(\frac{n}{d}\right). \end{aligned}$$

This proves that $\iota_0 \cdot \gamma = \iota_1$. By Möbius inversion, this implies that $\gamma = \iota_1 \cdot \mu$, and we see that $\gamma = \varphi$. We summarize our discussion in Theorem 10.1.

Theorem 10.1. *The function φ is a multiplicative function satisfying the identities*

$$\sum_{d|n} \varphi\left(\frac{n}{d}\right) = \sum_{d|n} \varphi(d) = n$$

and

$$\varphi(n) = \sum_{d|n} d \mu\left(\frac{n}{d}\right) = n \sum_{d|n} \frac{\mu(d)}{d}.$$

Furthermore, $\varphi(n)$ is the number of integers t such that $(t, n) = 1$ and $1 \leq t \leq n$, and if

$$1 < n = \prod_j p_j^{\delta_j}, \quad \delta_j > 0,$$

then

$$\varphi(n) = n \prod_j \left(1 - \frac{1}{p_j}\right).$$

The Euler φ -function is the best known member of the class of *totient* functions, and is sometimes referred to as the “Euler totient”. To define a totient, we first need to point out that a function $f \in \mathcal{A}$ is called *completely*

multiplicative if $f(mn) = f(m)f(n)$ for all, $n \in \mathcal{Z}^+$. Clearly, if f is completely multiplicative, then $f \in \mathcal{M}$, but not conversely—for example, $\sigma \in \mathcal{M}$ but $\sigma(4) \neq \sigma(2)\sigma(2)$.

Notice that in Exercise 7-2 it was shown that not only is $\iota_s \in \mathcal{M}$, but in fact ι_s is completely multiplicative for every complex s .

Definition. A function $f \in \mathcal{A}$ is called a totient if and only if $f = g \cdot h'$, where g is completely multiplicative and h' is the inverse of a completely multiplicative function.

EXERCISES

10-1. Show that φ is a totient.

10-2. If f is a totient, then $f \in \mathcal{M}$.

10-3. Prove the identities

$$(a) \quad \varphi \cdot \tau = \iota_1 \cdot \iota_0,$$

$$(b) \quad \varphi \cdot \sigma_s = \iota_1 \cdot \iota_s,$$

$$(c) \quad \varphi' = \iota_1' \cdot \iota_0.$$

Write these identities using the summation notation.

10-4. Prove or give a counter example for the following:

(a) If f is completely multiplicative, then so is f' .

(b) If f and g are completely multiplicative, then so is $f \cdot g$.

(c) If f is a totient, then so is f' .

(d) If f and g are totients, then so is $f \cdot g$.

10-5. Let $n \in \mathcal{Z}^+$, $k \in \mathcal{Z}$. Prove that the set

$$\{t: k < t \leq k + n, \quad (t, n) = 1\}$$

contains exactly $\varphi(n)$ members.

10-6. Find necessary and sufficient conditions that the characteristic function of a set $\mathcal{S} \subset \mathcal{Z}^+$ be completely multiplicative.

10-7. Find all solutions of the equation $\varphi(n) = 10$, $n \in \mathcal{Z}^+$.

10-8. Find $\varphi \cdot \varphi(n)$ for every $n \in \mathcal{Z}^+$. [Hint: Since the function $\varphi \cdot \varphi \in \mathcal{M}$, it suffices to find $\varphi \cdot \varphi(p^\beta)$ for every prime p , every $\beta \in \mathcal{Z}^+$.]

10-9. If $k \in \mathcal{Z}^+$, define the arithmetic function J_k to be the product of ι_k and μ , i.e., $J_k = \iota_k \cdot \mu$. The function J_k is called the *Jordan totient*; prove that J_k is in fact a totient for every $k \in \mathcal{Z}^+$. Notice that $J_1 = \varphi$, so that the Jordan totient is a generalization of the Euler totient. Prove that $J_k(1) = 1$, and if

$$1 < n = \prod_j p_j^{\beta_j},$$

then

$$J_k(n) = n^k \prod_j \left(1 - \frac{1}{p_j^k}\right).$$

10-10. If r, s are complex numbers, define $F_{r,s} = \iota_r \cdot \iota'_s$.

- (a) Prove that $F_{r,s}$ is a totient.
- (b) Show that the Jordan totient J_k is a special case of this totient $F_{r,s}$.
- (c) Use the results of Exercise 8-4 to prove that if

$$n = \prod_j p^{\beta_j},$$

$$F_{r,s}(n) = n^r \prod_j \left(1 - \frac{p^s}{p^r} \right).$$

10-11. If p is a prime but $2p + 1$ is not, then show that $\varphi(x) = 2p$ has no solutions.

10-12. If $d|n$ and $k \in \mathcal{Z}^+$, prove that

$$\varphi(nd^k) = d^k \varphi(n).$$

10-13. If $(m,n) = d$, show that

$$\varphi(mn) \varphi(d) = d \varphi(m) \varphi(n).$$

10-14. Prove that $\varphi(n) \tau(n) \geq n$, $n \in \mathcal{Z}^+$.

Chapter 3

CONGRUENCES AND RESIDUES

11. Complete Residue Systems

If m is a fixed positive integer, the division algorithm, with m as divisor, leads to an interesting classification of all integers into disjoint classes. If $a_1, a_2 \in \mathcal{L}$, we find the unique q_i, r_i ($i = 1, 2$) such that

$$(1) \quad \begin{aligned} a_1 &= q_1m + r_1, & 0 \leq r_1 < m \\ a_2 &= q_2m + r_2, & 0 \leq r_2 < m. \end{aligned}$$

Then we will write $a_1 \equiv a_2 \pmod{m}$, read “ a_1 is congruent to a_2 modulo m ,” if and only if $r_1 = r_2$; also $a_1 \not\equiv a_2 \pmod{m}$ means a_1 is not congruent to a_2 modulo m , and r_1 is called the least non-negative *residue* of a_1 modulo m . The first theorem gives a property equivalent to our defining property.

Note. Throughout this chapter we will assume that the modulus (usually to be denoted by m) is larger than 1.

Theorem 11.1. *A necessary and sufficient condition that $a_1 \equiv a_2 \pmod{m}$ is that $m|(a_1 - a_2)$.*

Proof. Suppose that the division algorithm gives the system of equations (1) and suppose $a_1 \equiv a_2 \pmod{m}$. Then $r_1 = r_2$ implies that $a_1 - q_1m = a_2 - q_2m$, and $m|(a_1 - a_2)$. Conversely, if $m|(a_1 - a_2)$, then for some b ,

$$bm = a_1 - a_2 = (q_1 - q_2)m + r_1 - r_2,$$

from which it follows that $m|(r_1 - r_2)$. But from the inequalities in (1) we have $-m < r_1 - r_2 < m$, and the only multiple of m in this range is $0 = r_1 - r_2$; thus the proof is complete.

The relation “congruence” on the set $\mathcal{L}$ of integers is an example of an *equivalence relation* on a set. In general, the relation $\mathcal{R}$ on a set $\mathcal{S}$ is an

equivalence relation if

- (a) $\mathcal{R}$ is reflexive, i.e., $a\mathcal{R}a$ for all $a \in \mathcal{S}$;
- (b) $\mathcal{R}$ is symmetric, i.e., if $a\mathcal{R}b$, then $b\mathcal{R}a$;
- (c) $\mathcal{R}$ is transitive, i.e., if $a\mathcal{R}b$ and $b\mathcal{R}c$, then $a\mathcal{R}c$.

It is well known that if $\mathcal{S}$ is a set and $\mathcal{R}$ is a given equivalence relation on $\mathcal{S}$, then $\mathcal{R}$ induces a partition of $\mathcal{S}$, and that, conversely, if $\mathcal{F}$ is any partition of $\mathcal{S}$, then $\mathcal{F}$ may be used to define an equivalence relation on $\mathcal{S}$. More precisely, we have the following.

Theorem 11.2. *Suppose $\mathcal{S}$ is a set and $\mathcal{R}$ is an equivalence relation on $\mathcal{S}$. For every $a \in \mathcal{S}$, the subset $\mathcal{S}_a$ is defined by*

$$\mathcal{S}_a = \{b \in \mathcal{S} : b\mathcal{R}a\}.$$

Then the collection $\mathcal{F} = \{\mathcal{S}_a : a \in \mathcal{S}\}$ is a partition of $\mathcal{S}$. Conversely, if $\mathcal{F}$ is any partition of a set $\mathcal{S}$, then the relation $\mathcal{R}$ defined by “if $a, b \in \mathcal{S}$, then $a\mathcal{R}b$ if and only if there exists a set $\mathcal{T} \in \mathcal{F}$ such that $a \in \mathcal{T}$ and $b \in \mathcal{T}$ ” is an equivalence relation on $\mathcal{S}$.

Proof. Suppose $\mathcal{R}$ is an equivalence relation on $\mathcal{S}$. Evidently

$$\bigcup_{\mathcal{S}_a \in \mathcal{F}} \mathcal{S}_a = \bigcup_{a \in \mathcal{S}} \mathcal{S}_a = \mathcal{S}.$$

so every element of $\mathcal{S}$ is in some subset $\mathcal{S}_a$. Assume that $\mathcal{S}_a, \mathcal{S}_b \in \mathcal{F}$ and $\mathcal{S}_a \cap \mathcal{S}_b \neq \emptyset$. Then there is some $c \in \mathcal{S}$ such that $c \in \mathcal{S}_a \cap \mathcal{S}_b$, and $c \in \mathcal{S}_a$ implies $c\mathcal{R}a$ and $c \in \mathcal{S}_b$ implies $c\mathcal{R}b$. Therefore, $a\mathcal{R}c$ by symmetry; together with $c\mathcal{R}b$, transitivity gives $a\mathcal{R}b$. Again by symmetry, we also have $b\mathcal{R}a$.

Let $x \in \mathcal{S}_a$. Then $x\mathcal{R}a$ and $a\mathcal{R}b$ imply $x\mathcal{R}b$, so $x \in \mathcal{S}_b$ and $\mathcal{S}_a \subset \mathcal{S}_b$. Similarly, if $y \in \mathcal{S}_b$, then $y\mathcal{R}b$ and $b\mathcal{R}a$ imply $y \in \mathcal{S}_a$, so $\mathcal{S}_b \subset \mathcal{S}_a$, hence $\mathcal{S}_a = \mathcal{S}_b$. Contrapositively, any two distinct subsets are disjoint. Therefore, $\mathcal{F}$ is a partition.

Conversely, assume $\mathcal{F}$ is any partition of $\mathcal{S}$ and consider $\mathcal{R}$ defined as in the statement of the theorem. For every $a \in \mathcal{S}$, a is in some subset, and therefore $a\mathcal{R}a$. Clearly $a\mathcal{R}b$ implies a and b are in the same subset, so $b\mathcal{R}a$. Finally, if $a\mathcal{R}b$ and $b\mathcal{R}c$, then a, b, c are in the same subset, so $a\mathcal{R}c$, and $\mathcal{R}$ is an equivalence relation. This completes the proof of the theorem.

The subsets $\mathcal{S}_a$ are called *equivalence classes*, and any element in $\mathcal{S}_a$ is called a *representative* of the class. When an operation, say $\times$, has been defined on $\mathcal{S}$, then it may be possible to define an analogous operation on the set $\mathcal{F}$ of equivalence classes: if $a\mathcal{R}b$ and $c\mathcal{R}d$ imply that $(a \times c)\mathcal{R}(b \times d)$, then we can define

$$\mathcal{S}_a \times \mathcal{S}_c = \mathcal{S}_{a \times c}.$$

In Theorem 11.3 we will show in effect that it is possible to define $+$ and $\cdot$ on the equivalence classes $\mathcal{S}_a$ induced by the congruence relation.

In the case of the congruence (mod m) relation on $\mathcal{Z}$, we agree to denote an equivalence class $\mathcal{Z}_a$ by any one of its representatives; thus the operations on the equivalence classes (also called *residue classes*) are defined so that the sum (or product) of the class containing $a \in \mathcal{Z}$ and the class containing $b \in \mathcal{Z}$ is the class containing $a + b$ (or ab). The resulting system is denoted $\mathcal{Z}/(m) = (\mathcal{Z}/(m), +, \cdot)$, and is called the system of *integers (mod m)*.

For example, take $m = 5$. The residue classes are

$$\mathcal{Z}_0 = \{\dots, -10, -5, 0, 5, 10, \dots\} = \mathcal{Z}_{-5} = \mathcal{Z}_5 = \dots$$

$$\mathcal{Z}_1 = \{\dots, -9, -4, 1, 6, 11, \dots\}$$

$$\mathcal{Z}_2 = \{\dots, -8, -3, 2, 7, 12, \dots\}$$

$$\mathcal{Z}_3 = \{\dots, -7, -2, 3, 8, 13, \dots\}$$

$$\mathcal{Z}_4 = \{\dots, -6, -1, 4, 9, 14, \dots\}.$$

Notice that $\mathcal{Z} = \mathcal{Z}_0 \cup \mathcal{Z}_1 \cup \dots \cup \mathcal{Z}_4$ and if $i \neq j$ ($0 \leq i, j \leq 4$), then $\mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset$. We denote the class $\mathcal{Z}_i$ by any of its representatives; for instance, -24 or -4 or 6 or 26 may be used to denote $\mathcal{Z}_1$. Instead of writing $\mathcal{Z}_2 + \mathcal{Z}_3 = \mathcal{Z}_{2+3} = \mathcal{Z}_5 = \mathcal{Z}_0$, we write $2 + 3 \equiv 5 \equiv 0 \pmod{5}$, and instead of $\mathcal{Z}_2 \cdot \mathcal{Z}_3 = \mathcal{Z}_6 = \mathcal{Z}_1$, we write $2 \cdot 3 \equiv 6 \equiv 1 \pmod{5}$. The algebraic system $\mathcal{Z}/(5)$ is the set with five elements, say $\mathcal{Z}_0, \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \mathcal{Z}_4$ together with the operations of addition and multiplication.

Theorem 11.3. *If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then $a \pm c \equiv b \pm d \pmod{m}$ and $ac \equiv bd \pmod{m}$.*

Proof. If $m|(a - b)$ and $m|(c - d)$, then $m|(a \pm c - (b \pm d))$. To prove that $ac \equiv bd$, say $mA = a - b$ and $mD = c - d$. Then

$$m^2AD = ac - bd - b(c - d) - d(a - b),$$

from which it is clear that $m|(ac - bd)$.

Corollary 11.3a. *If $a \equiv b \pmod{m}$, then for every $k \in \mathcal{Z}$, $ak \equiv bk \pmod{m}$.*

Corollary 11.3b. *If $a \equiv b \pmod{m}$, then for every $k \in \mathcal{Z}^+$, $a^k \equiv b^k \pmod{m}$.*

Theorem 11.4. *Suppose $(k, m) = d$. Then $ka \equiv kb \pmod{m}$ if and only if $a \equiv b \pmod{m/d}$.*

Proof. If $(k, m) = d$, say $dM = m$ and $dK = k$. If $m|(ka - kb)$, we have $mA = k(a - b)$ for some A . It follows that $MA = k(a - b)$ and since $(M, K) = 1$, $M|(a - b)$. But $M = m/d$, so $a \equiv b \pmod{m/d}$. Conversely, if $a \equiv b \pmod{m}$, then $M|(a - b)$, so $MB = a - b$ for some B . We have

$$ka - kb = kMB = KdMB = mKB,$$

and $m|(ka - kb)$.

If a is any integer and $m > 1$ is given, it is clear from the division algorithm that there are at most m possible values for the remainder r , and hence at most m residue classes in $\mathcal{Z}/(m)$. On the other hand, there are at least m classes because no two of the numbers $0, 1, \dots, m-1$ are congruent (mod m), so there are always exactly m distinct residue classes. A set of m integers which contains exactly one representative of each class is called a *complete residue system modulo m* , abbreviated CRS (mod m). The set $\{1, 2, \dots, m\}$ is called the *least positive CRS (mod m)*, the set $\{0, 1, \dots, m-1\}$ is the *least non-negative CRS (mod m)*, and the set

$$\left\{ \left\lfloor \frac{-m+2}{2} \right\rfloor, \left\lfloor \frac{-m+2}{2} \right\rfloor + 1, \dots, 0, 1, \dots, \left\lfloor \frac{m}{2} \right\rfloor \right\}$$

is the *absolutely least CRS (mod m)*.

For example, if $m = 6$, the absolutely least CRS is the set $\{-2, -1, 0, 1, 2, 3\}$. If $m = 7$, it is $\{-3, -2, \dots, 2, 3\}$.

Theorem 11.5. *If $\{a_1, \dots, a_m\}$ is a CRS (mod m), and if $(k, m) = 1$, then $\{ka_1, \dots, ka_m\}$ is also a CRS (mod m).*

Proof. Suppose that for some $i \neq j$ we have

$$ka_i \equiv ka_j \pmod{m}.$$

Since $(k, m) = 1$, Theorem 11.4 implies $a_i \equiv a_j \pmod{m}$. But this contradicts the fact that the numbers $a_1, \dots, a_m$ form a CRS. Hence the set $\{ka_1, \dots, ka_m\}$ is a set of m pairwise incongruent elements, and so is a CRS (mod m).

EXERCISES

- 11-1. Prove that congruence modulo m is an equivalence relation on $\mathcal{Z}$.
- 11-2. Prove Corollary 11.3a and Corollary 11.3b.
- 11-3.* If $a \equiv b \pmod{m}$ and $0 < d|m$, prove $a \equiv b \pmod{d}$.
- 11-4. Find x in the least non-negative CRS (mod 7) such that
 - (a) $8^{25} \equiv x \pmod{7}$,
 - (b) $2^{30} \equiv x \pmod{7}$,
 - (c) $(-4)^{10} \equiv x \pmod{7}$.
- 11-5. Suppose $\mathcal{S} = (\mathcal{S}, +, \cdot)$ is an algebraic system and let $\mathcal{S}'$ denote the subset of $\mathcal{S}$ obtained by deleting from $\mathcal{S}$ the identity with respect to the operation "+". Then $\mathcal{S}$ is said to be a *commutative ring with identity* if $(\mathcal{S}, +)$ is a commutative group, $(\mathcal{S}', \cdot)$ is a commutative semigroup with identity, and " $\cdot$ " is distributive over "+", that is, for all $a, b, c \in \mathcal{S}$, $a \cdot (b + c) = a \cdot b + a \cdot c$. $\mathcal{S}$ is a *field* if both $(\mathcal{S}, +)$ and $(\mathcal{S}', \cdot)$ are commutative groups and the distributive law holds.
 - (a) Prove $(\mathcal{Z}/(m), +, \cdot)$ is a commutative ring with identity.

- (b) If m is prime, prove $\mathcal{Z}/(m)$ is a field.
 (c) Prove that if $\mathcal{Z}/(m)$ is a field, then m is prime.
 11-6. Show that the set

$$\mathcal{S} = \left\{ \left[\frac{-m+2}{2} \right], \left[\frac{-m+2}{2} \right] + 1, \dots, 0, 1, \dots, \left[\frac{m}{2} \right] \right\}$$

contains exactly m distinct elements, no two of which are congruent modulo m . Thus prove that the absolutely least CRS (mod m) is a CRS (mod m).

- 11-7. Give an example to show that for some a, b, k, m we may have $ka \equiv kb \pmod{m}$, but $a \not\equiv b \pmod{m}$.
 11-8. Give an example to show that Theorem 11.5 may not be true if $(k, m) \neq 1$. Prove or disprove the following statement: If $\{a_1, \dots, a_m\}$ is a CRS (mod m), and if $(k, m) > 1$, then $\{ka_1, \dots, ka_m\}$ is *not* a CRS (mod m).
 11-9.* If $b \equiv c \pmod{m_1}$ and $b \equiv c \pmod{m_2}$, prove $b \equiv c \pmod{\langle m_1, m_2 \rangle}$.

12. Linear Congruences

We will find necessary and sufficient conditions for the existence of a solution x of the congruence $ax + c \equiv 0 \pmod{m}$, and show that this is related to the problem of finding solutions x, y of the linear Diophantine equation $ax + by = n$. (By a Diophantine equation we shall mean an equation, in one or more unknowns, whose solutions are required to be integers.) Later in this section we will show how to solve a system of linear congruences.

If $f(x)$ is a polynomial with integral coefficients, we say x_0 is a solution of the congruence $f(x) \equiv 0 \pmod{m}$ if $f(x_0) \equiv 0 \pmod{m}$. The problem of finding solutions of a congruence is very different from finding solutions of polynomial equations. For example, the equation $x^3 - 3x^2 + 2x = 0$ has three solutions, $x = 0, 1, 2$, but the congruence

$$x^3 - 3x^2 + 2x \equiv 0 \pmod{3}$$

is satisfied by every integer. As another example, $x - 2 = 0$ has a unique solution, but $x - 2 \equiv 0 \pmod{5}$ is solved by every $x \equiv 2 \pmod{5}$. Finally, $x^2 - 2 = 0$ has two roots, but the congruence $x^2 - 2 \equiv 0 \pmod{3}$ has no solution. These examples show that a polynomial congruence may have no solutions or infinitely many solutions. That there are no other possibilities is implied by our next theorem.

Theorem 12.1. Suppose $f(x)$ is a polynomial with integral coefficients. If $a \equiv b \pmod{m}$, then $f(a) \equiv f(b) \pmod{m}$.

Proof. Let

$$f(x) = \sum_{k=0}^n c_k x^k.$$

From Corollary 11.3b, we know that $a \equiv b \pmod{m}$ implies $a^k \equiv b^k \pmod{m}$ for $k = 1, \dots, n$, and from Corollary 11.3a, that $c_k a^k \equiv c_k b^k$ for $k = 1, \dots, n$. Therefore,

$$\sum_{k=0}^n c_k a^k \equiv \sum_{k=0}^n c_k b^k \pmod{m};$$

but this says $f(a) \equiv f(b) \pmod{m}$.

Corollary 12.1. *If a is a solution of $f(x) \equiv 0 \pmod{m}$, then for every b such that $b \equiv a \pmod{m}$, b is also a solution.*

Because of this last result, we agree to say a congruence has n solutions if it has exactly n incongruent solutions. If m is small, solutions can be found by testing each number in a CRS $\pmod{m}$, but for large m this method is impractical. For a linear congruence one may always use the Euclidean algorithm to find a solution (if one exists), and we will show that it is possible to get all solutions from a given one. Moreover, we can easily tell how many solutions will exist. The problem of finding solutions of a congruence is somewhat easier if we first prove this result on solutions of a linear Diophantine equation.

Theorem 12.2. *The equation $ax + by = n$ has a solution x, y if and only if $d = (a, b) \mid n$. If there is a solution x_0, y_0 , then there are infinitely many solutions: for every $t \in \mathcal{Z}$, then x, y of the form $x = x_0 + (b/d)t, y = y_0 + (-a/d)t$ is a solution, and every solution x, y is of this form for some $t \in \mathcal{Z}$.*

Proof. Since the first part of this theorem is merely a restatement of Corollary 2.1a, we proceed with the proof of the second part. Suppose that x_0, y_0 is a solution. For every $t \in \mathcal{Z}$,

$$a\left(x_0 + \frac{b}{d}t\right) + b\left(y_0 - \frac{a}{d}t\right) = ax_0 + by_0 = n,$$

so $x_0 + (b/d)t, y_0 + (a/d)(-t)$ is a solution. If x, y is any solution, then we define X and Y by $X = x - x_0, Y = y - y_0$. Then

$$\begin{aligned} n = ax + by &= a(X + x_0) + b(Y + y_0) \\ &= ax_0 + by_0 + aX + bY. \end{aligned}$$

This gives $aX + bY = 0$, from which we have

$$(2) \quad \frac{a}{d}X + \frac{b}{d}Y = 0.$$

Therefore, $(b/d)|(a/d)X$, but $(a/d, b/d) = 1$, so $(b/d)|X$; say $(b/d)t = X$. Then (2) implies $Y = -(a/d)t$. Thus $x = x_0 + X = x_0 + (b/d)t$ and $y = y_0 - (a/d)t$ as required.

Corollary 12.2. *The linear congruence $ax \equiv n \pmod{m}$ has solutions if and only if $d = (a, m)|n$. If there is a solution, then there are exactly d (incongruent) solutions. In particular, if $(a, m) = 1$, then the congruence has a unique solution.*

Proof. The first part is evident since there exists an x such that $ax \equiv n \pmod{m}$ if and only if there exist x and y such that $ax + my = n$. If there is a solution x_0 of the congruence, then all solutions x are of the form $x = x_0 + (m/d)t$, $t \in \mathcal{Z}$. If

$$x_1 = x_0 + \frac{m}{d}t_1 \quad \text{and} \quad x_2 = x_0 + \frac{m}{d}t_2$$

are solutions, and if $x_1 \equiv x_2 \pmod{m}$, then

$$x_0 + \frac{m}{d}t_1 \equiv x_0 + \frac{m}{d}t_2 \pmod{m}$$

$$t_1 \equiv t_2 \left(\text{mod } \frac{m}{(m, m/d)} \right)$$

$$t_1 \equiv t_2 \pmod{d}.$$

Therefore, two solutions, x_1 and x_2 , are incongruent $\pmod{m}$ if and only if t_1 and t_2 are incongruent $\pmod{d}$, so there are d distinct ways to obtain incongruent solutions.

Example. Consider the congruence $3x \equiv 9 \pmod{12}$. By Theorem 11.4, this is equivalent to $x \equiv 3 \pmod{4}$, which obviously has the solution $x = 3$. Then the original congruence will have $(3, 12) = 3$ solutions, all of the form $3 + (12/3)t$. Therefore, $x = 3, 7, 11$ are the solutions in the least positive CRS $\pmod{12}$ and $x = -5, -1, 3$ are the solutions in the absolutely least CRS.

The following result gives a method for finding a solution of a system of linear congruences under certain conditions.

Theorem 12.3. *(Chinese Remainder Theorem.) If $(m_i, m_j) = 1$ for $1 \leq i < j \leq r$, then the system*

$$(3) \quad \begin{cases} x \equiv a_1 \pmod{m_1} \\ x \equiv a_2 \pmod{m_2} \\ \dots \\ x \equiv a_r \pmod{m_r} \end{cases}$$

always has a solution x which is unique modulo

$$\prod_{i=1}^r m_i.$$

Proof. Let

$$M = \prod_{i=1}^r m_i$$

and $M = A_i m_i$, $i = 1, \dots, r$. For each i , $(A_i, m_i) = 1$. Let y_i be the unique $(\text{mod } m_i)$ solution of

$$A_i y_i \equiv 1 \pmod{m_i}.$$

Notice that if $i \neq j$, $A_j \equiv 0 \pmod{m_i}$, so that

$$x = \sum_{i=1}^r a_i A_i y_i$$

is a solution of (3).

Suppose that x and z are any two solutions of (3). For every i , $x \equiv a_i \equiv z \pmod{m_i}$ implies that $m_i | (x - z)$ and hence, $M | (x - z)$ by Exercise 2-1(b). That is, any two solutions are congruent $(\text{mod } M)$.

Example. Solve

$$(4) \quad \begin{cases} 2x \equiv 1 \pmod{3} \\ 3x \equiv 2 \pmod{4} \\ 4x \equiv 2 \pmod{5} \end{cases}$$

Evidently we must have

$$(5) \quad \begin{cases} x \equiv 2 \pmod{3} \\ x \equiv 2 \pmod{4} \\ x \equiv 3 \pmod{5} \end{cases}$$

Now we use the Chinese Remainder Theorem to construct the solution of (5). We find $M = 60$, $A_1 = 20$, $A_2 = 15$, $A_3 = 12$, $y_1 = 2$, $y_2 = y_3 = 3$. Then $\sum a_i A_i y_i = 278 \equiv 38 \pmod{60}$. Thus, $x = 38$ is the solution of both (4) and (5).

EXERCISES

12-1. Find all solutions x, y of $2x + 3y = 7$ such that $-10 < x < 10$.

12-2. For the congruence $12x \equiv 16 \pmod{28}$, find all solutions in

- (a) the least positive CRS $(\text{mod } 28)$,
- (b) the absolutely least CRS $(\text{mod } 28)$.

- 12-3. Solve the system $x \equiv 1 \pmod{3}$, $x \equiv 2 \pmod{5}$, $x \equiv 3 \pmod{7}$.
 12-4. Solve the system $3x \equiv 1 \pmod{4}$, $2x \equiv 4 \pmod{6}$, $2x \equiv 3 \pmod{11}$.
 12-5. Sketch the graph of the equation $4x + 6y = 10$. On this line, call a point P a "solution point" if P has abscissa x , ordinate y , and x, y is a solution of the Diophantine equation $4x + 6y = 10$. Find a solution point which is
 (a) closer to the origin than every other solution point;
 (b) closer to the y -axis than every other solution point;
 (c) at least as close to the x -axis as every other solution point.
 12-6. Consider the graph of $ax + by = n$, where $a, b, n \in \mathcal{Z}$. Prove that the set of solution points (see Exercise 12-5) is just the set of points on the line $ax + by = n$ which have integer coordinates.

13. Reduced Residue Systems

If $\mathcal{S}$ is any CRS $(\text{mod } m)$, the subset of $\mathcal{S}$ which consists of those $a \in \mathcal{S}$ such that $(a, m) = 1$ is called a *reduced residue system modulo m* , abbreviated RRS $(\text{mod } m)$. As with a CRS, we may speak about the least positive or the absolutely least RRS $(\text{mod } m)$. Evidently the set $\{a: 1 \leq a \leq m, (a, m) = 1\}$ is the least positive RRS $(\text{mod } m)$, and the set contains exactly $\varphi(m)$ elements. If $\mathcal{R}$ is any other RRS $(\text{mod } m)$, then $\mathcal{R}$ also contains $\varphi(m)$ elements and every element of $\mathcal{R}$ is congruent $(\text{mod } m)$ to exactly one element in the least positive RRS. We can see this is so as follows. Suppose $b \in \mathcal{R}$, and consider the congruence $x \equiv b \pmod{m}$. This has a unique solution x_0 in the least positive CRS $(\text{mod } m)$. But since $x_0 - b = km$ for some k , if $(x_0, m) = d > 1$, then $d|(m, b)$, which contradicts that b is in a RRS $(\text{mod } m)$. Therefore, $(x_0, m) = 1$ and x_0 is in the least positive RRS.

Lemma 13.1. *If $\{a_1, \dots, a_{\varphi(m)}\}$ is a RRS $(\text{mod } m)$ and if $(k, m) = 1$, then $\{ka_1, \dots, ka_{\varphi(m)}\}$ also is a RRS $(\text{mod } m)$.*

The proof of the lemma is the same as the proof of Theorem 11.5. The lemma leads to this important result.

Theorem 13.1. (Euler's Theorem.) *If $(k, m) = 1$, then*

$$k^{\varphi(m)} \equiv 1 \pmod{m}.$$

Proof. Let $\{a_1, \dots, a_{\varphi(m)}\}$ be a RRS $(\text{mod } m)$. Then also $\{ka_1, \dots, ka_{\varphi(m)}\}$ is a RRS $(\text{mod } m)$, so that in some order the a_i are congruent $(\text{mod } m)$ to the ka_j . Therefore,

$$\begin{aligned} \prod_{i=1}^{\varphi(m)} (ka_i) &\equiv \prod_{i=1}^{\varphi(m)} a_i \pmod{m}; \\ k^{\varphi(m)} \prod a_i &\equiv \prod a_i \pmod{m}. \end{aligned}$$

But $(a_i, m) = 1$ for $i = 1, \dots, \varphi(m)$ and $(\prod a_i, m) = 1$ so we may cancel this common factor from the last congruence without affecting the modulus.

A special case of Euler's Theorem was first proved by Fermat.

Corollary 13.1. (*Fermat's Theorem.*) *If p is a prime, then*

$$a^p \equiv a \pmod{p}.$$

The Fermat (as well as the Euler) Theorem may be used to simplify computations. For example, if asked to show that $23^{475} \equiv 2 \pmod{7}$, we would notice that $23^{475} \equiv 2^{475} \equiv 2^{6(79)+1} \equiv (2^6)^{79} \cdot 2 \equiv 1^{79} \cdot 2 \equiv 2 \pmod{7}$.

Suppose now that for some k , $k^{\varphi(m)} \equiv 1 \pmod{m}$. Then the congruence $kx \equiv 1 \pmod{m}$ has a solution $x = k^{\varphi(m)-1}$, so by Corollary 12.2, $(k, m) \mid 1$, or k and m are relatively prime. This shows that the converse of the Euler Theorem is true, namely, if $k^{\varphi(m)} \equiv 1 \pmod{m}$, then $(k, m) = 1$.

Our next theorem will be applied in Section 14 in the study of another arithmetic function.

Theorem 13.2. *Suppose $(m, n) = 1$. Let $\mathcal{R}_1$ be a RRS $\pmod{m}$ and $\mathcal{R}_2$ be a RRS $\pmod{n}$. Then the set*

$$\mathcal{R} = \{km + jn : k \in \mathcal{R}_2, j \in \mathcal{R}_1\}$$

is a RRS $\pmod{mn}$.

Proof. There are $v\mathcal{R}_2 = \varphi(n)$ choices for k and $v\mathcal{R}_1 = \varphi(m)$ choices for j , so at most $\varphi(m)\varphi(n) = \varphi(mn)$ elements in $\mathcal{R}$. That there are exactly $\varphi(mn)$ elements may be shown by an indirect argument. Suppose for some $j_1, j_2 \in \mathcal{R}_1$ and some $k_1, k_2 \in \mathcal{R}_2$, $k_1 \leq k_2$, we have

$$k_1m + j_1n = k_2m + j_2n.$$

Then

$$(6) \quad (j_1 - j_2)n = (k_2 - k_1)m,$$

so $n \mid (k_2 - k_1)m$. But since the k_i are from a RRS $\pmod{n}$, this implies $k_1 = k_2$. Then (6) yields $j_1 = j_2$. Thus, $v\mathcal{R} = \varphi(mn)$.

We now show that no two elements of $\mathcal{R}$ are congruent $\pmod{mn}$. Suppose instead that for k_i, j_i ($i = 1, 2$) as above, we have

$$k_1m + j_1n - (k_2m + j_2n) = qmn$$

for some q . Then $(k_1 - k_2)m = (qm - j_1 + j_2)n$. But, as above, $n \mid (k_1 - k_2)m$ implies $k_1 = k_2$ and it follows that $j_1 = j_2$. This shows that $\mathcal{R}$ is contained in some CRS $\pmod{mn}$.

Finally, assume $km + jn \in \mathcal{R}$, and say $d = (km + jn, mn)$. Now d divides any linear combination of these numbers, and since $km^2 = m(km + jn) - j(mn)$, $d \mid km^2$. Write $d = ab$, where $a \mid m$, $b \mid n$. Since $b \mid d$, also $b \mid km^2$. But $(b, m) = 1$,

so $b|k$, and $b|n$, therefore $b|(k,n) = 1$. In a similar manner, $a = 1$, so $d = 1$. Thus, $\mathcal{R}$ consists of $\varphi(mn)$ elements from a CRS $(\bmod mn)$, each of which is relatively prime to mn , so $\mathcal{R}$ is a RRS $(\bmod mn)$.

EXERCISES

- 13-1. Give a proof of Fermat's Theorem.
 13-2. In $\mathcal{Z}/(m)$ consider the subset of elements $\mathcal{Z}_i$ with $(i,m) = 1$. Prove that this subset is a group under multiplication. [This result may be stated roughly by saying that a RRS $(\bmod m)$ is a group under multiplication if one recognizes that the product of two elements is equivalent to any congruent integer.]
 13-3.* Suppose for some $k \in \mathcal{Z}$, $n \in \mathcal{Z}^+$ we have

$$(7) \quad k^n \equiv 1 \pmod{m}.$$

Prove that $(k,m) = 1$. If n is the smallest such integer for which (7) holds, prove that $n|\varphi(m)$. [Hint: Use the division algorithm on n and $\varphi(m)$.]

- 13-4. Suppose $(m,n) = 1$. Let $\mathcal{C}_1$ be a CRS $(\bmod m)$ and $\mathcal{C}_2$ be a CRS $(\bmod n)$. Prove that the set

$$\mathcal{C} = \{km + jn : k \in \mathcal{C}_2, j \in \mathcal{C}_1\}$$

is a CRS $(\bmod mn)$.

- 13-5. If $(a,m) = 1$, show that the solution of $ax \equiv n \pmod{m}$ is

$$x = a^{\varphi(m)-1}n.$$

- 13-6. If $\{a_1, \dots, a_{\varphi(m)}\}$ is a RRS $(\bmod m)$ and if $(k,m) > 1$, prove that $\mathcal{K} = \{ka_1, \dots, ka_{\varphi(m)}\}$ is not a RRS $(\bmod m)$. Does there exist a non-empty subset of $\mathcal{K}$ which is contained in some RRS $(\bmod m)$?
 13-7. Suppose $\{a_1, \dots, a_{\varphi(m)}\}$ is a RRS $(\bmod m)$. Use the division algorithm to find q_j, r_j [$j = 1, \dots, \varphi(m)$] such that $a_j = mq_j + r_j$, $0 \leq r_j < m$. Prove that $\{r_1, \dots, r_{\varphi(m)}\}$ is the least positive RRS $(\bmod m)$.
 13-8. Using $\{1,3\}$ and $\{1,2,3,4\}$ for the RRS $(\bmod 4)$ and $(\bmod 5)$, respectively, compute the corresponding RRS $(\bmod 20)$ obtained by the method of Theorem 13.2. Notice that though we started with the least positive RRS $(\bmod 4)$ and the least positive RRS $(\bmod 5)$, we do not get the least positive RRS $(\bmod 20)$.

14. Ramanujan's Trigonometric Sum

Trigonometric functions are encountered frequently in number theory. As a typical example we consider Ramanujan's trigonometric sum. We assume the reader is familiar with the following properties of the exponential

function of a complex variable:

$$(8) \quad e^{z_1 + z_2} = e^{z_1} e^{z_2} \quad \text{and} \quad (e^{z_1})^{z_2} = e^{z_1 z_2};$$

$$(9) \quad \text{if } z \text{ is real, } e^{iz} = \cos z + i \sin z.$$

It will be convenient here to let $\mathcal{R}_n$ denote the least positive RRS (mod n).

Definition. If $s \in \mathcal{L}^+$, the function $c_s \in \mathcal{A}$ is defined by

$$c_s(n) = \sum_{k \in \mathcal{R}_n} e^{2\pi i s k / n};$$

the function c_s is called Ramanujan's trigonometric sum.

We will first show that if the numbers k in the summation are from any RRS (mod n), the value $c_s(n)$ is not changed.

Lemma 14.1. If $\mathcal{T}$ is any RRS (mod n), then

$$c_s(n) = \sum_{t \in \mathcal{T}} e^{2\pi i s t / n}.$$

Proof. We know that if $\mathcal{T}$ is any RRS (mod n) and $\mathcal{R}_n$ is the least positive RRS (mod n), then for each $t \in \mathcal{T}$, there is a unique $k \in \mathcal{R}_n$ such that $t \equiv k \pmod{n}$. Thus, for every $t \in \mathcal{T}$, there is a unique k such that

$$t = k + n q(k) \text{ for some integer } q(k).$$

Therefore,

$$\begin{aligned} \sum_{t \in \mathcal{T}} e^{2\pi i s t / n} &= \sum_{k \in \mathcal{R}_n} e^{2\pi i s (k + n q(k)) / n} \\ &= \sum_{k \in \mathcal{R}_n} e^{2\pi i s k / n} e^{2\pi i s q(k)} \\ &= c_s(n) \end{aligned}$$

because, by (9), for every k , $e^{2\pi i s q(k)} = 1$.

We can now prove the following theorem which makes it clear that c_s is a real-valued function.

Theorem 14.1.

$$c_s(n) = \sum_{k \in \mathcal{R}_n} \cos \frac{2\pi s k}{n}.$$

Proof. Take $\mathcal{T} = \{-k : k \in \mathcal{R}_n\}$. [That $\mathcal{T}$ is a RRS (mod n) is left as an exercise.] By Lemma 14.1,

$$\begin{aligned} \sum_{k \in \mathcal{R}_n} e^{2\pi i s k / n} &= \sum_{k \in \mathcal{R}_n} e^{2\pi i s (-k) / n} \\ \sum_k \left\{ \cos \frac{2\pi s k}{n} + i \sin \frac{2\pi s k}{n} \right\} &= \sum_k \left\{ \cos \frac{2\pi s (-k)}{n} + i \sin \frac{2\pi s (-k)}{n} \right\} \end{aligned}$$

$$\sum_k \cos \frac{2\pi sk}{n} + i \sum_k \sin \frac{2\pi sk}{n} = \sum_k \cos \frac{2\pi sk}{n} - i \sum_k \sin \frac{2\pi sk}{n}.$$

In the last equation we have used the familiar facts that $\cos(-x) = \cos x$ and $\sin(-x) = -\sin x$ for all real x . From the equation above it is clear that

$$(10) \quad \sum_{k \in \mathcal{R}_n} \sin \frac{2\pi sk}{n} = 0.$$

Applying (9) to the definition of $c_s(n)$ and using (10), the theorem is proved.

Corollary 14.1. *If $n|s$, then $c_s(n) = \varphi(n)$.*

Proof. Say $nd = s$. Then

$$c_s(n) = \sum_{k \in \mathcal{R}_n} \cos 2\pi dk = \sum_{k \in \mathcal{R}_n} 1 = \varphi(n).$$

Before giving an evaluation of $c_s(n)$ it is helpful to prove that c_s is multiplicative.

Theorem 14.2. *For every s , $c_s \in \mathcal{M}$.*

Proof. That $c_s(1) = 1$ follows from the Corollary 14.1. Assume $(m, n) = 1$. Then

$$\begin{aligned} c_s(m)c_s(n) &= \sum_{j \in \mathcal{R}_m} e^{2\pi i s j/m} \sum_{k \in \mathcal{R}_n} e^{2\pi i s k/n} \\ &= \sum_{\substack{j \in \mathcal{R}_m \\ k \in \mathcal{R}_n}} e^{2\pi i s (j/m + k/n)} \\ &= \sum_{\substack{j \in \mathcal{R}_m \\ k \in \mathcal{R}_n}} e^{2\pi i s (jn + km)/mn}, \end{aligned}$$

but by Theorem 13.2, this is just $c_s(mn)$, so $c_s \in \mathcal{M}$.

Taking the sum of exponentials over somewhat larger ranges than the type we have been considering, we get a geometric series. For example,

$$\sum_{k=1}^n e^{2\pi i s k/n} = \begin{cases} e^{2\pi i s/n} \frac{\{e^{2\pi i s/n}\}^n - 1}{e^{2\pi i s/n} - 1} & \text{if } e^{2\pi i s/n} \neq 1, \\ n & \text{if } e^{2\pi i s/n} = 1. \end{cases}$$

But $e^{2\pi i s/n} = \cos(2\pi s/n) + i \sin(2\pi s/n) = 1$ if and only if $n|s$. From (8) and (9), $\{e^{2\pi i s/n}\}^n = e^{2\pi i s} = 1$. Thus, we have

$$\sum_{k=1}^n e^{2\pi i s k/n} = \begin{cases} 0, & n \nmid s, \\ n, & n|s. \end{cases}$$

Furthermore, if n is a prime power, $n = p^\beta$, then

$$\mathcal{R}_{p^\beta} = \{k : 1 \leq k \leq p^\beta\} - \{tp : 1 \leq t \leq p^{\beta-1}\}.$$

That is, the numbers in $\mathcal{R}_{p^\beta}$ are just the numbers from 1 to p^β whose gcd with p^β is not larger than 1; the numbers whose gcd with p^β is larger than 1 are all of the form tp , $1 \leq t \leq p^{\beta-1}$.

We write $p^\beta \parallel s$ (read “ p^β is an exact divisor of s ”) to mean $p^\beta | s$ and $p^{\beta+1} \nmid s$. Then

$$\begin{aligned} c_s(p^\beta) &= \sum_{k \in \mathcal{R}_{p^\beta}} e^{2\pi i sk/p^\beta} \\ &= \sum_{k=1}^{p^\beta} e^{2\pi i sk/p^\beta} - \sum_{t=1}^{p^{\beta-1}} e^{2\pi i t p/p^\beta} \\ &= \begin{cases} 0, & p^\beta \nmid s \\ p^\beta, & p^\beta | s \end{cases} - \begin{cases} 0, & p^{\beta-1} \nmid s \\ p^{\beta-1}, & p^{\beta-1} | s \end{cases} \\ &= \begin{cases} 0, & p^{\beta-1} \nmid s \\ -p^{\beta-1}, & p^{\beta-1} \parallel s \\ p^\beta - p^{\beta-1}, & p^\beta | s. \end{cases} \end{aligned}$$

It is now easy to prove Theorem 14.3.

Theorem 14.3.

$$\sum_{d|n} c_s(d) = \begin{cases} n, & n|s \\ 0, & n \nmid s \end{cases}$$

Proof. It suffices to prove this theorem for $n = p^\beta$, prime p . If $p^\beta | s$, then $c_s(d) = \varphi(d)$ for every $d|p^\beta$ and from Theorem 10.1 it follows that $\sum c_s(d) = p^\beta$. If $p^\beta \nmid s$, say $p^\alpha \parallel s$, $0 \leq \alpha < \beta$, and

$$\begin{aligned} \sum_{d|p^\beta} c_s(d) &= c_s(1) + \sum_{j=1}^{\alpha} c_s(p^j) + \sum_{j=\alpha+1}^{\beta} c_s(p^j) \\ &= 1 + \sum_{j=1}^{\alpha} (p^j - p^{j-1}) + (-p^{(\alpha+1)-1}) \\ &= 0. \end{aligned}$$

Corollary 14.3.

$$c_s(n) = \sum_{\substack{d|n \\ d|s}} \mu\left(\frac{n}{d}\right) d$$

Proof. Let $\delta_s \in \mathcal{A}$ be defined by $\delta_s(n) = 0$ if $n \nmid s$, $= n$ if $n \mid s$. Theorem 14.3 says $c_s \cdot \iota_0(n) = \delta_s(n)$. By Möbius inversion, $c_s(n) = \mu \cdot \delta_s(n)$. Therefore,

$$c_s(n) = \sum_{d \mid n} \mu\left(\frac{n}{d}\right) \delta_s(d) = \sum_{\substack{d \mid n \\ d \mid s}} \mu\left(\frac{n}{d}\right) d.$$

EXERCISES

14-1. Find $c_1(n)$ for every $n \in \mathcal{Z}^+$. (*Hint*: Use Corollary 14.3.)

14-2. Prove $\delta_s \in \mathcal{M}$.

14-3. Prove $c_n \cdot \iota'_1(n) = \mu(n)$ for every $n \in \mathcal{Z}^+$.

14-4. Prove the set $\mathcal{T}$ (defined in the proof of Theorem 14.1) is a RRS (mod n).

14-5. Prove that

$$c_s(n) = \sum_{d \mid (n, s)} \mu\left(\frac{n}{d}\right) d.$$

14-6. Is c_s completely multiplicative? Is δ_s completely multiplicative?

14-7. Prove

$$(a) \quad c_s \cdot \tau = \delta_s \cdot \iota_0;$$

$$(b) \quad c_s \cdot \sigma = \delta_s \cdot \iota_1.$$

15. Wilson's Theorem

In this section we will prove the result known as Wilson's Theorem: The integer $n > 1$ is prime if and only if $(n-1)! \equiv -1 \pmod{n}$. Though this is not a practical test for primality for large values of n , the result is of theoretical interest and arises as a natural consequence of a theorem of Lagrange, which is proved below.

Theorem 15.1. (*Lagrange.*) Suppose p is a prime and

$$f(x) = \sum_{k=0}^n c_k x^k$$

is a polynomial with integral coefficients. Then either the congruence $f(x) \equiv 0 \pmod{p}$ has at most n incongruent solutions or $c_k \equiv 0 \pmod{p}$ for $k = 0, 1, \dots, n$.

Proof. We prove this theorem by induction on the degree of $f(x)$. If $n = 1$ and $c_1 \not\equiv 0 \pmod{p}$, then there is just one solution of $f(x) \equiv 0 \pmod{p}$, by Corollary 12.2. Suppose Lagrange's Theorem is true for polynomials of degree $\leq n-1$; assume that $f(x)$ is of degree n , and the congruence $f(x) \equiv 0 \pmod{p}$ has at least $n+1$ solutions $a_1, \dots, a_n, a_{n+1}$ which are incongruent (mod p).

If $c_n \equiv 0 \pmod{p}$, we construct the polynomial

$$g(x) = \sum_{k=0}^{n-1} c_k x^k;$$

clearly all $n+1$ solutions of $f(x) \equiv 0 \pmod{p}$ are also solutions of $g(x) \equiv 0 \pmod{p}$, and since $g(x)$ is of degree $n-1$, we have contradicted the inductive supposition. Therefore, $c_n \not\equiv 0 \pmod{p}$.

Dividing $f(x)$ by $x - a_{n+1}$, we find

$$f(x) = (x - a_{n+1}) q(x) + r$$

where $r \in \mathcal{Z}$ and $q(x)$ is a polynomial with integral coefficients, of degree $n-1$, and with leading coefficient c_n . Since

$$0 \equiv f(a_{n+1}) = (a_{n+1} - a_{n+1}) q(a_{n+1}) + r \equiv r \pmod{p},$$

we have $r \equiv 0$ and $f(x) \equiv (x - a_{n+1})q(x) \pmod{p}$; that is, the polynomial $f(x)$ and the polynomial $(x - a_{n+1})q(x)$ have corresponding coefficients which are congruent $\pmod{p}$. For every $k \neq n+1$, $0 \equiv f(a_k) \equiv (a_k - a_{n+1})q(a_k) \pmod{p}$; since p is prime, and $a_k - a_{n+1} \not\equiv 0$, we have by Theorem 4.1 that $q(a_k) \equiv 0 \pmod{p}$ for $k = 1, \dots, n$. Therefore, $q(x) \equiv 0 \pmod{p}$ is a polynomial congruence of degree $n-1$ and has at least n incongruent solutions, which is a contradiction. This concludes the proof.

Notice that the theorem of Lagrange is not true if the modulus is not prime. For example, the congruence $x^2 - 5x \equiv 0 \pmod{6}$ has four solutions, $x = 0, 2, 3, 5$. Also the number of solutions may be strictly less than the degree, as in $x^3 - 2 \equiv 0 \pmod{5}$, which has only one solution, $x = 3$.

Theorem 15.2. (Wilson.) *A necessary and sufficient condition that $n > 1$ be prime is that $(n-1)! \equiv -1 \pmod{n}$.*

Proof. Suppose n is an odd prime (if $n = 2$, the theorem is obvious). Consider the two polynomials

$$f(x) = x^{n-1} - 1$$

$$g(x) = (x-1)(x-2)\cdots(x-(n-1))$$

and define $h(x) = f(x) - g(x)$. Clearly $h(x)$ has degree $n-2$. The congruence $g(x) \equiv 0 \pmod{n}$ is satisfied by each of the numbers $1, 2, \dots, n-1$, and by Euler's Theorem these numbers are also solutions of $f(x) \equiv 0 \pmod{n}$. Therefore $h(x) \equiv 0 \pmod{n}$ has at least $n-1$ solutions, so the coefficients are all $0 \pmod{n}$. In particular, the constant term in $h(x)$ is $0 \pmod{n}$, or

$$-1 - (-1)^{n-1}(n-1)! \equiv 0 \pmod{n},$$

and this proves the necessity of the condition.

Conversely, if n is composite it has a prime divisor p , and p is a factor of $(n-1)!$, or $(n-1)! \equiv 0 \pmod{p}$. Evidently, then, by Exercise 11-3, $(n-1)! \not\equiv -1 \pmod{n}$.

EXERCISES

- 15-1. Prove that if n is composite, then $(n-1)! \equiv 0 \pmod{n}$ except for $n = 4$.
- 15-2. Consider again the polynomial $h(x)$ defined in the proof of Wilson's Theorem. Let b_k denote the coefficient of x^k in $h(x)$, for $k = 1, 2, \dots, n-2$. We proved that $b_k \equiv 0 \pmod{n}$. Describe the numbers b_k .
- 15-3. If $f(x) = g(x)h(x)$, prove that every solution of $f(x) \equiv 0 \pmod{p}$, prime p , is either a solution of $g(x) \equiv 0 \pmod{p}$ or of $h(x) \equiv 0 \pmod{p}$. Show by an example that this need not be true if the modulus is composite.
- 15-4. Suppose $f(x) = g^n(x)$, $n \in \mathcal{Z}^+$. If p is a prime, prove that every solution of $f(x) \equiv 0 \pmod{p^n}$ is a solution of $g(x) \equiv 0 \pmod{p}$, and vice versa.
- 15-5. Prove that $[(n-1)!/n]$ is even for $n > 4$.
- 15-6. Show that Fermat's Theorem and Wilson's Theorem together imply and are implied by the following statement: For all $a \in \mathcal{Z}$ and all primes p , $a^p + a(p-1)!$ is a multiple of p .
- 15-7. Show that the expression $a^p + a(p-1)!$ in Exercise 15-6 may be replaced by $a + a^p(p-1)!$.

16. Primitive Roots

Definition. If $(a, m) = 1$, then a is said to belong to (the exponent) $t \pmod{m}$ if and only if t is the least positive integer x such that $a^x \equiv 1 \pmod{m}$.

Clearly, if a belongs to $t \pmod{m}$, then $1 \leq t \leq \phi(m)$. In fact, t is a divisor of $\phi(m)$ (see Exercise 13-3). More generally, suppose a belongs to $t \pmod{m}$. Then $a^x \equiv 1 \pmod{m}$ if and only if $t|x$.

Theorem 16.1. If a belongs to $t_1 \pmod{m_1}$ and a belongs to $t_2 \pmod{m_2}$, then a belongs to $\langle t_1, t_2 \rangle \pmod{\langle m_1, m_2 \rangle}$.

Proof. Say a belongs to $t \pmod{\langle m_1, m_2 \rangle}$ and let $T = \langle t_1, t_2 \rangle$. Now $a^t \equiv 1 \pmod{\langle m_1, m_2 \rangle}$ implies that $a^t \equiv 1 \pmod{m_1}$ and $t_1|t$. Similarly, $t_2|t$, so $T|t$. But $a^T \equiv 1 \pmod{m_1}$ and $a^T \equiv 1 \pmod{m_2}$ imply that $a^T \equiv 1 \pmod{\langle m_1, m_2 \rangle}$ by Exercise 11-9, so $t|T$.

Theorem 16.2. If a belongs to $t \pmod{m}$, then a^u belongs to $t/(t, u) \pmod{m}$.

Proof. Suppose a^u belongs to $s \pmod{m}$, and let $d = (t, u)$. Since a belongs to t , $a^{us} \equiv 1$ implies $t|us$ and $(t/d)|s$. Similarly, since a^u belongs to s , $a^{u(t/d)} \equiv 1$ implies $s|(t/d)$.

Definition. If a belongs to $\phi(m) \pmod{m}$, then a is called a primitive root of m .

Example. It is easy to see that 2 is a primitive root of 5, because $2^2 \equiv 4$, $2^3 \equiv 3$, $2^4 \equiv 1 \pmod{5}$ and $\phi(5) = 4$. Also, 3 belongs to 4 $\pmod{5}$ so 3 is also a primitive root of 5. However, 1 belongs to 1 $\pmod{5}$ and 4 belongs to 2 $\pmod{5}$, so neither 1 nor 4 is a primitive root of 5.

Example. Not every $m > 1$ has primitive roots. For instance, consider $m = 8$. Evidently, if a is even, then $a^x \equiv 1 \pmod{8}$ has no solution by Exercise 13-3, so no even integer can be a primitive root of 8. If a is odd, then either $a \equiv \pm 1$ or $a \equiv \pm 3 \pmod{8}$; in any case, $a^2 \equiv 1 \pmod{8}$, so every odd integer belongs to 2 $\pmod{8}$ and hence is not a primitive root, since $\phi(8) = 4$.

In view of the last two examples it is natural to ask what values of m have primitive roots. We will now concern ourselves with proving that the complete answer to this question is: The integer $m > 1$ has a primitive root if and only if $m = 2, 4, p^\beta$, or $2p^\beta$, where p is an odd prime.

Theorem 16.3. Suppose p is an odd prime and $t|(p-1)$. Then there are exactly $\phi(t)$ incongruent numbers which belong to $t \pmod{p}$.

Proof. Let $\psi(t)$ be the number of incongruent numbers which belong to $t \pmod{p}$. We recall that if there is an a which belongs to $t \pmod{p}$, then $t|\phi(p)$; contrapositively, if $t \nmid (p-1)$, then $\psi(t) = 0$. Since there are exactly $p-1$ incongruent numbers which are prime to p , and each of these belongs to some unique exponent $\pmod{p}$, we have

$$(11) \quad p-1 = \sum_{t|(p-1)} \psi(t).$$

For each $t|(p-1)$, $\psi(t) \geq 0$. If $\psi(t) \neq 0$, there is some a which belongs to $t \pmod{p}$. Consider the congruence $x^t \equiv 1 \pmod{p}$. Evidently each of the incongruent integers $a, a^2, \dots, a^t$ is a solution, and by Lagrange's Theorem there can be no others. Hence every number which belongs to t is of the form a^u , $1 \leq u \leq t$. But by Theorem 16.2, a^u belongs to $t/(u, t)$, so a^u belongs to t if and only if $(u, t) = 1$, $1 \leq u \leq t$. There are clearly $\phi(t)$ such numbers. Hence, for every $t|(p-1)$, either $\psi(t) = 0$ or $\psi(t) = \phi(t)$. From (11) and Theorem 10.1, we get

$$p-1 = \sum_{t|(p-1)} \psi(t) \leq \sum_{t|(p-1)} \phi(t) = p-1.$$

Therefore for every $t|(p-1)$, $\psi(t) = \phi(t)$, and the proof is complete.

Corollary 16.3. If p is an odd prime, there are $\phi(\phi(p)) = \phi(p-1)$ primitive roots of p . In particular, since $\phi(p-1) \geq 1$, every odd prime has a primitive root.

We now introduce a function $\lambda \in \mathcal{A}$ and show that for every a prime to m , $a^{\lambda(m)} \equiv 1 \pmod{m}$, that is, $\lambda(m)$ is a *universal exponent* for m . This will tell us that if $\lambda(m) < \phi(m)$, then m cannot have primitive roots.

$$\lambda(m) = \begin{cases} \phi(m) & \text{if } m = 1, 2, 4, p^\beta, 2p^\beta \text{ (} p \text{ odd prime);} \\ \frac{1}{2}\phi(m) & \text{if } m = 2^\beta, \beta \geq 3; \\ \langle \lambda(2^\beta), \lambda(p_1^{\beta_1}), \dots, \lambda(p_r^{\beta_r}) \rangle & \text{if } m = 2^\beta \prod_{i=1}^r p_i^{\beta_i}, p_i \text{ odd primes.} \end{cases}$$

Theorem 16.4. For every $(a, m) = 1$, $a^{\lambda(m)} \equiv 1 \pmod{m}$.

Proof. If $m = 2, 4, p^\beta$, or $2p^\beta$ the result is immediate from Euler's Theorem. Suppose $m = 2^\beta$, $\beta \geq 3$. If $(a, 2^\beta) = 1$, then a is odd and we have already observed that $a^2 \equiv 1 \pmod{8}$, or that

$$a^{\lambda(2^3)} \equiv 1 \pmod{2^3}.$$

Suppose we have shown that for $3 \leq \beta - 1$,

$$a^{\lambda(2^{\beta-1})} \equiv 1 \pmod{2^{\beta-1}}.$$

Then for some k ,

$$a^{\lambda(2^{\beta-1})} = 1 + k2^{\beta-1}.$$

Squaring both sides, we have

$$a^{2\lambda(2^{\beta-1})} = 1 + k2^\beta + k^2 2^{2\beta-2}$$

so that $a^{\lambda(2^\beta)} \equiv 1 \pmod{2^\beta}$.

Finally, suppose that

$$m = 2^\beta \prod_{i=1}^r p_i^{\beta_i}.$$

If $(a, m) = 1$, then $1 = (a, 2^\beta) = (a, p_1^{\beta_1}) = \dots = (a, p_r^{\beta_r})$, and we have already shown that

$$a^{\lambda(2^\beta)} \equiv 1 \pmod{2^\beta}$$

$$a^{\lambda(p_1^{\beta_1})} \equiv 1 \pmod{p_1^{\beta_1}}$$

...

$$a^{\lambda(p_r^{\beta_r})} \equiv 1 \pmod{p_r^{\beta_r}}.$$

But $\lambda(2^\beta) \parallel \lambda(m)$, and for $i = 1, \dots, r$, $\lambda(p_i^{\beta_i}) \parallel \lambda(m)$, so

$$a^{\lambda(m)} \equiv 1 \pmod{2^\beta}$$

$$a^{\lambda(m)} \equiv 1 \pmod{p_1^{\beta_1}}$$

...

$$a^{\lambda(m)} \equiv 1 \pmod{p_r^{\beta_r}}.$$

Hence, each of the above moduli is a divisor of $a^{\lambda(m)} - 1$, and since the moduli are relatively prime in pairs, their product m is a divisor of $a^{\lambda(m)} - 1$. This completes the proof.

Corollary 16.4. *If m is a multiple of any number of the type (a) $4p$, where p is an odd prime, or (b) pq , where $p \neq q$ are odd primes, or (c) 8 , then m cannot have a primitive root.*

Proof. Suppose

$$m = 2^\beta \prod_{i=1}^r p_i^{\beta_i}.$$

If m has a divisor of type (a), ($\beta \geq 2$ and $r \geq 1$), or of type (b), ($r \geq 2$), then

$$\begin{aligned} \lambda(m) &= \langle \lambda(2^\beta), \dots, \lambda(p_r^{\beta_r}) \rangle \\ &< \lambda(2^\beta) \cdots \lambda(p_r^{\beta_r}) \\ &\leq \varphi(2^\beta) \cdots \varphi(p_r^{\beta_r}) = \varphi(m). \end{aligned}$$

If $8|m$ ($\beta \geq 3$), again $\lambda(m) < \varphi(m)$. In any case, no a prime to m can belong to $\varphi(m) \pmod{m}$.

Lemma 16.5. *If a is a primitive root of p^β , then either a belongs to $\varphi(p^\beta) \pmod{p^{\beta+1}}$ or a is a primitive root of $p^{\beta+1}$. Every primitive root of $p^{\beta+1}$ is a primitive root of p^β .*

Proof. Suppose a belongs to $t \pmod{p^{\beta+1}}$. Then $t|\varphi(p^{\beta+1})$. From $a^t \equiv 1 \pmod{p^{\beta+1}}$ we have the same congruence $\pmod{p^\beta}$, so $\varphi(p^\beta)|t$, since a belongs to $\varphi(p^\beta) \pmod{p^\beta}$. Therefore, either $t = \varphi(p^\beta)$ or $t = p\varphi(p^\beta) = \varphi(p^{\beta+1})$.

If a is a primitive root of $p^{\beta+1}$, and if a belongs to $t \pmod{p^\beta}$, then $t|\varphi(p^\beta)$. Since $a^t \equiv 1 \pmod{p^\beta}$, we have $a^t = 1 + kp^\beta$. Raising both sides of this equation to the power p , we get $a^{pt} \equiv 1 \pmod{p^{\beta+1}}$. Therefore, $\varphi(p^{\beta+1})|pt$, so $t = \varphi(p^\beta)$ and a is a primitive root of p^β .

Theorem 16.5. *For every $\beta \in \mathcal{X}^+$, there exists a primitive root of p^β , where p is an odd prime. If a is a primitive root of p , then at least one of a or $a + p$ is a primitive root of p^2 . If a is a primitive root of p^β ($\beta \geq 2$), then a is a primitive root of $p^{\beta+1}$. Consequently, there exists a number a (independent of β) such that a is a primitive root of p^β for all $\beta \in \mathcal{X}^+$.*

Proof. We know by Corollary 16.3 that there is a primitive root a of p . By Lemma 16.5, either a is a primitive root of p^2 or a belongs to $\varphi(p) \pmod{p^2}$. In the latter case, consider $a + p$, which is also a primitive root of p .

$$\begin{aligned} (a + p)^{\varphi(p)} &= a^{\varphi(p)} + (p - 1)a^{\varphi(p)-1}p + \cdots \\ &\equiv 1 - pa^{\varphi(p)-1} \pmod{p^2}. \end{aligned}$$

Now if $a + p$ belongs to $\varphi(p)$, then $-pa^{\varphi(p)-1} \equiv 0 \pmod{p^2}$, or $a^{\varphi(p)-1} \equiv 0 \pmod{p}$, which is impossible. Therefore, $a + p$ belongs to $\varphi(p^2) \pmod{p^2}$, by Lemma 16.5.

Assume we have proved the existence of a primitive root for every p^x , $x = 1, \dots, \beta$ ($\beta \geq 2$). If a is a primitive root of p^β , then by the lemma a is also a primitive root of $p^{\beta-1}$, so $a^{\varphi(p^{\beta-1})} = 1 + kp^{\beta-1}$, and $p \nmid k$. Raising both sides to the p^{th} power we get

$$a^{p\varphi(p^{\beta-1})} \equiv 1 + kp^\beta \pmod{p^{\beta+1}}.$$

Since $p \nmid k$,

$$a^{p\varphi(p^{\beta-1})} = a^{\varphi(p^\beta)} \not\equiv 1 \pmod{p^{\beta+1}},$$

so by the lemma a is a primitive root of $p^{\beta+1}$. The induction is complete.

Corollary 16.5. *For every $\beta \in \mathcal{Z}^+$, there exists a primitive root of $2p^\beta$.*

Proof. We know there is a primitive root a of p^β . If q is even, then $a + p^\beta$ is an odd primitive root of p^β , so we may assume a is odd. Hence a is also a primitive root of 2. By Theorem 16.1, a belongs to $\langle \varphi(2), \varphi(p^\beta) \rangle = \varphi(2p^\beta) \pmod{2p^\beta}$.

In those cases where m has primitive roots, namely for $m = 2, 4, p^\beta$, or $2p^\beta$, we can state exactly how many there are.

Theorem 16.6. *If m has a primitive root, then m has exactly $\varphi(\varphi(m))$ incongruent primitive roots.*

Proof. Suppose a is a primitive root of m . Then we can easily see (Exercise 16-4) that

$$\mathcal{S} = \{a^k : k = 1, 2, \dots, \varphi(m)\}$$

is a RRS $\pmod{m}$. Consider the congruence

$$(12) \quad x^{\varphi(m)} \equiv 1 \pmod{m}.$$

Every $a^k \in \mathcal{S}$ is a solution of (12). We know by Exercise 13-3 that all the solutions of (12) are in $\mathcal{S}$. By Theorem 16.2, the number of a^k , $1 \leq k \leq \varphi(m)$, which belong to $\varphi(m)$ is just the number of k , $1 \leq k \leq \varphi(m)$, such that $(k, \varphi(m)) = 1$. There are $\varphi(\varphi(m))$ such k .

We saw in Theorem 16.5 that every primitive root of p^β , $\beta \geq 2$, is also a primitive root of $p^{\beta+1}$. This does not hold for $\beta = 1$; for example, 8 is a primitive root of 3, but not of 3^2 . In fact, for every odd prime p , there exists a number g such that g is a primitive root of p , but not of p^2 . This can be seen easily from the last theorem. Let $g_1, \dots, g_r$, $r = \varphi(\varphi(p))$, be the incongruent $\pmod{p}$ primitive roots of p . Then all the numbers $g_i + kp$ ($i = 1, \dots, r$; $k = 0, 1, \dots, p-1$) are primitive roots of p and are incongruent $\pmod{p^2}$. But there are rp such numbers, and $rp > \varphi(p^2)$, so not all of them can be primitive roots of p^2 . This proves the following.

Corollary 16.6. *For every odd prime p there are $\phi(p-1) = p \phi(\phi(p)) - \phi(\phi(p^2)) = p \phi(p-1) - \phi(p^2 - p)$ numbers incongruent (mod p^2) which are primitive roots of p , but not of p^2 .*

EXERCISES

- 16-1. Find at least one primitive root for each of the following moduli: 7, 9, 13, 15, 17, 25, 27.
- 16-2. Suppose $(a, m) = 1$ and consider the sequence $\{b_1, b_2, b_3, \dots\}$ where b_i is the least positive residue (mod m) of a^i . Prove this sequence is periodic with period equal to the exponent to which a belongs (mod m).
- 16-3. Suppose a belongs to $t_1 \pmod{m_1}$ and a belongs to $t_2 \pmod{m_2}$. To what exponent (mod $\langle m_1, m_2 \rangle$) does a^v belong?
- 16-4.* If a is a primitive root of m , prove $\{a, a^2, \dots, a^{\phi(m)}\}$ is a RRS (mod m).
- 16-5. Prove that if a is a primitive root of $p^{\beta+1}$ ($\beta \geq 2$), then so is $a + kp^{\beta}$ for every k .
- 16-6. If $\beta \geq 2$ and $\mathcal{G}$ is a set of all the incongruent primitive roots of p^{β} , where p is an odd prime, prove that the set

$$\mathcal{H} = \{a + kp^{\beta} : a \in \mathcal{G}, k = 0, 1, \dots, p-1\}$$

is a set of all the incongruent primitive roots of $p^{\beta+1}$.

- 16-7. Let p be an odd prime and let $(a, p) = 1$. Consider the sequence $\{t_1, t_2, t_3, \dots\}$ where a belongs to $t_x \pmod{p^x}$. Describe the sequence $\{t_x\}$.
- 16-8. Find an integer which is a primitive root of both 29 and 73, using the information that 2 is a primitive root of 29 and 5 is a primitive root of 73.

17. Quadratic Residues

We will now study polynomial congruences of the type

$$(13) \quad f(y) = Ay^2 + By + C \equiv 0 \pmod{p}, (A, p) = 1,$$

Since some special techniques are required to deal with the prime 2, *throughout this section p will always denote an odd prime*. We first observe that it suffices to be able to solve congruences of the type

$$(14) \quad x^2 \equiv a \pmod{p}$$

in order to find all the solutions of (13).

In (14), take $a \equiv B^2 - 4AC \pmod{p}$. If x_1 is a solution of (14), then there corresponds a unique y_1 such that $2Ay_1 \equiv x_1 - B$, since $(2A, p) = 1$. Now we have

$$0 \equiv x_1^2 - a \equiv (2Ay_1 + B)^2 - (B^2 - 4AC) \equiv 4Af(y_1) \pmod{p}.$$

But $(4A, p) = 1$, so we have found a solution $y = y_1$ of (13). Conversely, if there are no solutions of (14) for $a \equiv B^2 - 4AC$, then there can be no solution of (13).

Definition. Suppose $(a, p) = 1$. We say a is a quadratic residue of p if the congruence (14) is solvable. If (14) is not solvable, we say a is a quadratic non-residue of p .

If x is a solution of (14), then $p - x$ is also a solution and $x \not\equiv p - x \pmod{p}$, so by Lagrange's Theorem these are all the solutions. Consider the numbers $1^2, 2^2, \dots, \{(p-1)/2\}^2$ which are obviously all quadratic residues of p . These are all distinct $\pmod{p}$, because if $1 \leq a < b \leq (p-1)/2$ and $a^2 \equiv b^2 \pmod{p}$, then $(a-b)(a+b) = a^2 - b^2 \equiv 0 \pmod{p}$ and either $p|(a-b)$ or $p|(a+b)$; but both are impossible. If $(p-1)/2 < a \leq p-1$, then $a^2 \equiv (p-a)^2$ and $1 \leq p-a \leq (p-1)/2$. Thus all the quadratic residues of p are the distinct numbers $1^2, 2^2, \dots, \{(p-1)/2\}^2$. This proves Theorem 17.1.

Theorem 17.1. There are exactly $(p-1)/2$ quadratic residues and $(p-1)/2$ quadratic non-residues of p .

We will show next that a is a residue or a non-residue according as $a^{(p-1)/2}$ is congruent to 1 or $-1 \pmod{p}$. The statement of the result is facilitated by defining the Legendre symbol.

Definition. Suppose $(a, p) = 1$. The Legendre symbol (a/p) is defined to be $+1$ if a is a quadratic residue of p , and -1 if a is a non-residue of p .

Theorem 17.2. (Euler's criterion.) If $(a, p) = 1$, then $a^{(p-1)/2} \equiv (a/p) \pmod{p}$. Hence, a is a residue of p if and only if $a^{(p-1)/2} \equiv 1 \pmod{p}$.

Proof. If $(a, p) = 1$, then

$$0 \equiv a^{p-1} - 1 \equiv (a^{(p-1)/2} - 1)(a^{(p-1)/2} + 1) \pmod{p},$$

so $a^{(p-1)/2} \equiv \pm 1 \pmod{p}$.

Suppose $x^2 \equiv a \pmod{p}$ has a solution $x = x_1$. Then

$$a^{(p-1)/2} \equiv (x_1^2)^{(p-1)/2} \equiv 1 \pmod{p}.$$

Contrapositively, if $a^{(p-1)/2} \equiv -1$, then a is a non-residue.

Conversely, suppose $a^{(p-1)/2} \equiv 1 \pmod{p}$, and let g be a primitive root of p . We know (Exercise 16-4) that the powers of g form a RRS $\pmod{p}$, so there is some j , $1 \leq j \leq p-1$, such that $g^j \equiv a \pmod{p}$. Then

$$g^{j(p-1)/2} \equiv a^{(p-1)/2} \equiv 1 \pmod{p}$$

so $\varphi(p) = (p-1) \mid \{j(p-1)/2\}$. Therefore j is even, say $j = 2J$. But then $g^j \equiv (g^J)^2 \equiv a \pmod{p}$, so that $(a/p) = 1$.

Corollary 17.2. Suppose $(a, p) = 1$ and $(b, p) = 1$. Then $(a^2/p) = 1$, $(ab/p) = (a/p)(b/p)$, and if $a \equiv b \pmod{p}$, then $(a/p) = (b/p)$.

Note. The equation $(ab/p) = (a/p)(b/p)$ says that the product of two residues is a residue, the product of two non-residues is a residue, and the product of a residue and a non-residue is a non-residue.

The problem of deciding if a given integer is a residue or non-residue of p is made easier by the properties of the Legendre symbol given in Corollary 17.2. In fact, it suffices to know (q/p) for all primes q in order to know (a/p) for all a . If $a > 1$ and $(a, p) = 1$, suppose $a = \prod q_i^{\beta_i}$. Then $(a/p) = \prod (q_i/p)^{\beta_i}$. Obviously, this has the same value as the simplified form

$$(a/p) = \prod_{\beta_i \text{ odd}} (q_i/p).$$

If $a < -1$, then

$$(a/p) = (-1)^{(p-1)/2} \prod_{\beta_i \text{ odd}} (q_i/p).$$

Example. Find $(-540/7)$. Since $-540 = -2^2 3^3 5$,

$$(-540/7) = (-1)^3 (2/7)^2 (3/7)^3 (5/7) = -(3/7)(5/7).$$

But 1, 2, 4 are the residues of 7, so $(-540/7) = -1$. More simply, since $-540 \equiv -1 \pmod{7}$, $(-540/7) = (-1/7) = (-1)^3 = -1$.

To be able to find (q/p) for prime q , we will use the following result.

Theorem 17.3. (Gauss Lemma.) Suppose $(a, p) = 1$. Let u denote the number of the integers $a, 2a, \dots, \{(p-1)/2\}a$ whose absolutely least residues $(\text{mod } p)$ are negative. Then $(a/p) = (-1)^u$.

Proof. Let $r_1, \dots, r_v$ be those residues such that $0 < r_i < p/2$ and $-s_1, \dots, -s_u$ be those residues such that $-p/2 < -s_j < 0$. Obviously no two r_i are equal and no two s_j are equal. Also, no r_i can equal any s_j , because if so,

$$ma \equiv r_i = s_j \equiv -na \pmod{p}$$

for some m, n between 0 and $p/2$; but then $a(m+n) \equiv 0 \pmod{p}$, which is impossible since $a \not\equiv 0$ and $0 < m+n < p$. Therefore, the numbers $r_1, \dots, r_v, s_1, \dots, s_u$ are just the integers $1, 2, \dots, (p-1)/2$ in some order, and

$$a(2a) \cdots \left(\frac{p-1}{2} a \right) \equiv r_1 \cdots r_v (-s_1) \cdots (-s_u) \equiv (-1)^u 1 \cdot 2 \cdots \frac{p-1}{2} \pmod{p}$$

$$a^{(p-1)/2} \left(\frac{p-1}{2} \right)! \equiv (-1)^u \left(\frac{p-1}{2} \right)! \pmod{p}.$$

Thus, $a^{(p-1)/2} \equiv (-1)^u \equiv (a/p) \pmod{p}$. But since $(-1)^u$ and (a/p) are ± 1 , we have $(-1)^u = (a/p)$.

The proof of our next lemma is geometric, and a few introductory remarks will explain the notions involved. A *lattice point* in a Cartesian plane is a point with integer coordinates. If $\mathcal{R}$ is a bounded region in the plane, we let $\Lambda_{\mathcal{R}}$ denote the number of lattice points in $\mathcal{R}$. If $\mathcal{R}$ and $\mathcal{S}$ are bounded plane regions, then

$$(15) \quad \Lambda_{\mathcal{R} \cup \mathcal{S}} = \Lambda_{\mathcal{R}} + \Lambda_{\mathcal{S}} - \Lambda_{\mathcal{R} \cap \mathcal{S}};$$

that is, the number of lattice points in $\mathcal{R} \cup \mathcal{S}$ is the number in $\mathcal{R}$ plus the number in $\mathcal{S}$ minus the number which have been counted twice (because they are both in $\mathcal{R}$ and in $\mathcal{S}$).

Recall that $[x]$ is the largest integer $\leq x$.

Lemma 17.4. *Suppose p and q are distinct odd primes, and define*

$$S(m, n) = \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} [jm/n].$$

Then $S(p, q) + S(q, p) = \frac{1}{2}(p-1) \cdot \frac{1}{2}(q-1)$.

Proof. For definiteness, assume $p > q$. In the plane, consider the three bounded regions $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ (see Figure 1) formed by the lines

$$L_1: y = \frac{1}{2}$$

$$L_2: y = \frac{q-1}{2}$$

$$L_3: y = \frac{q}{p}x$$

$$L_4: x = \frac{p-1}{2}$$

$$L_5: x = \frac{1}{2}.$$

As in the figure, let $\mathcal{A}$ denote the region bounded by L_i , $i = 2, 3, 5$; $\mathcal{B}$, the region bounded by L_i , $i = 1, 2, 3, 4$; and $\mathcal{C}$, the region bounded by L_i , $i = 2, 3, 4$.

Suppose j is an integer, $1 \leq j \leq (q-1)/2$, and consider the number of lattice points on the line $y = j$ and in $\mathcal{A}$. Clearly, these will be the points with coordinates $(1, j), (2, j), \dots, (\zeta, j)$, where ζ is the largest possible abscissa of any point on this line $y = j$ which is not to the right of L_3 . Since the lines

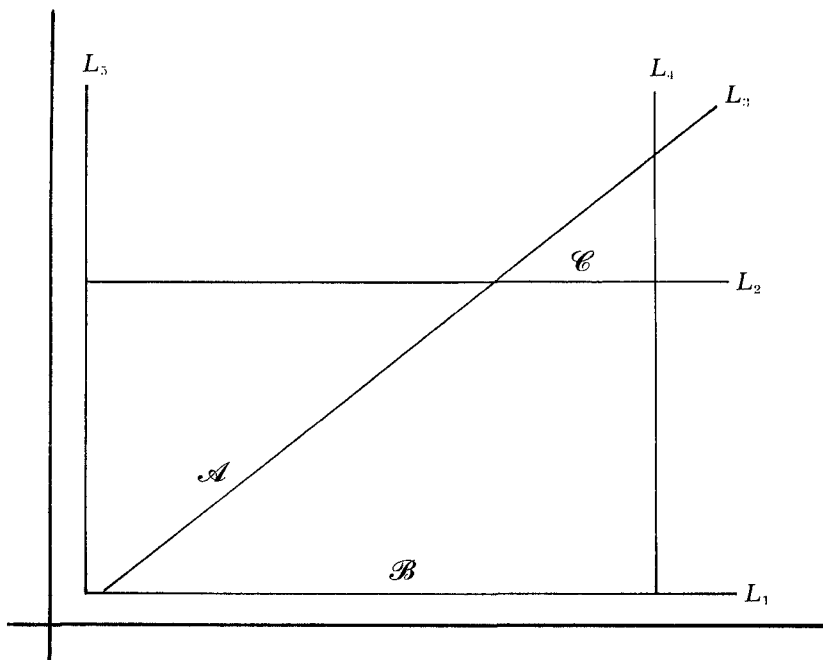


Figure 1

$y = j$ and L_3 intersect at $(jp/q, j)$, we see there are $\zeta = [jp/q]$ lattice points on $y = j$ and in $\mathcal{A}$. Consequently, the number $S(p, q)$ is just $\Lambda_{\mathcal{A}}$, and by similar arguments, $S(q, p) = \Lambda_{\mathcal{B} \cup \mathcal{C}}$, so

$$(16) \quad S(p, q) + S(q, p) = \Lambda_{\mathcal{A}} + \Lambda_{\mathcal{B} \cup \mathcal{C}}.$$

Now we note that the only lattice points in $\mathcal{C}$ are on L_2 . This may be seen by observing that

$$\left(\frac{p-1}{2}, \frac{q-1}{2} \right) = L_4 \cap L_2$$

is a lattice point and

$$\left(\frac{p-1}{2}, \frac{q(p-1)}{2p} \right) = L_3 \cap L_4$$

is less than 1 unit from $L_4 \cap L_2$. Therefore, every point in $\mathcal{C}$ is less than 1 unit from L_2 . This shows that $\Lambda_{\mathcal{C}} = \Lambda_{\mathcal{B} \cap \mathcal{C}}$, and from (15) we get $\Lambda_{\mathcal{B} \cup \mathcal{C}} = \Lambda_{\mathcal{B}}$.

It is also easy to see that there are no lattice points on the segment of L_3 which is between L_1 and L_2 , since p and q are distinct primes. Thus, $\Lambda_{\mathcal{A} \cap \mathcal{B}} = 0$,

so that $\Lambda_{\mathcal{A} \cup \mathcal{B}} = \Lambda_{\mathcal{A}} + \Lambda_{\mathcal{B}}$. Making substitutions in (16) we have

$$S(p, q) + S(q, p) = \Lambda_{\mathcal{A} \cup \mathcal{B}}.$$

But obviously, $\Lambda_{\mathcal{A} \cup \mathcal{B}}$, the number of pairs (x, y) of integers satisfying $\frac{1}{2} \leq x \leq \frac{1}{2}(p-1)$ and $\frac{1}{2} \leq y \leq \frac{1}{2}(q-1)$, is $\frac{1}{2}(p-1) \cdot \frac{1}{2}(q-1)$. This completes the proof.

Theorem 17.4. (*Quadratic reciprocity law.*) If p and q are distinct odd primes, then

$$(p/q)(q/p) = (-1)^{\frac{1}{2}(p-1) \cdot \frac{1}{2}(q-1)}.$$

Proof. Suppose $1 \leq k \leq \frac{1}{2}(p-1)$. Then, from the division algorithm, $kq = p[kq/p] + t_k$, $0 < t_k \leq p-1$. If $t_k < p/2$, then t_k is one of the r_i in the proof of Theorem 17.3, and if $p/2 < t_k$, then $t_k - p = -s_j$ for some j , $1 \leq j \leq u$. Recall that we showed in Theorem 17.3 that the r_i and the s_j are just the integers $1, 2, \dots, (p-1)/2$ in some order. Hence

$$\begin{aligned} \frac{p^2-1}{8} &= \sum_{k=1}^{(p-1)/2} k = \sum r_i + \sum s_j = \sum r_i + \sum_{t_k > p/2} (p - t_k) \\ &= \sum r_i + up - \sum_{t_k > p/2} t_k, \end{aligned}$$

or

$$(17) \quad \sum r_i = \frac{p^2-1}{8} - up + \sum_{t_k > p/2} t_k.$$

Also,

$$\begin{aligned} \frac{q(p^2-1)}{8} &= \sum_{k=1}^{(p-1)/2} kq = \sum_{k=1}^{(p-1)/2} \left\{ p \left[\frac{kq}{p} \right] + t_k \right\} \\ &= pS(q, p) + \sum r_i + \sum_{t_k > p/2} t_k. \end{aligned}$$

Combining this last equation with (17) we get

$$p(S(q, p) - u) = \frac{(q-1)(p^2-1)}{8} - 2 \sum_{t_k > p/2} t_k.$$

But $(p^2-1)/8$ is an integer and $q-1$ is even, so

$$p(S(q, p) - u) \equiv 0 \pmod{2}$$

$$S(q, p) \equiv u \pmod{2}.$$

Then $(-1)^{S(q, p)} = (-1)^u = (q/p)$. By a symmetric argument, $(-1)^{S(p, q)} = (p/q)$ and

$$(q/p)(p/q) = (-1)^{S(q, p) + S(p, q)} = (-1)^{\frac{1}{2}(p-1) \cdot \frac{1}{2}(q-1)}.$$

It is sometimes convenient to use the Law of Quadratic Reciprocity in the form

$$(p/q) = (-1)^{\frac{1}{2}(p-1) \cdot \frac{1}{2}(q-1)} (q/p).$$

To complete our knowledge we need to know the quadratic character of 2 modulo an odd prime. This is given by Theorem 17.5.

Theorem 17.5.

$$(2/p) = (-1)^{(p^2-1)/8} = \begin{cases} 1, & p \equiv \pm 1 \pmod{8} \\ -1, & p \equiv \pm 3 \pmod{8} \end{cases}$$

Proof. By the Gauss Lemma, $(2/p)$ will be $(-1)^u$, where u is the number of the integers $2, 2 \cdot 2, 3 \cdot 2, \dots, \frac{1}{2}(p-1) \cdot 2$ which are larger than $p/2$. For $1 \leq k \leq (p-1)/2$, $2k > p/2$ if and only if $p/4 < k \leq (p-1)/2$. The number of such k is $u = (p-1)/2 - [p/4]$.

We need only know if u is even or odd. If $p \equiv \pm 1 \pmod{8}$, then $u \equiv 0 \pmod{2}$, and if $p \equiv \pm 3 \pmod{8}$, then $u \equiv 1 \pmod{2}$. But also, if $p \equiv \pm 1 \pmod{8}$, then

$$\frac{p^2-1}{8} \equiv 0 \pmod{2},$$

and if $p \equiv \pm 3 \pmod{8}$, then

$$\frac{p^2-1}{8} \equiv 1 \pmod{2}.$$

Example. Find $(60/23)$. We have

$$\begin{aligned} (60/23) &= (14/23) = (2/23)(7/23) = (-1)^{(23^2-1)/8}(7/23) = (7/23) \\ &= (-1)^{\frac{1}{2}(23-1) \cdot \frac{1}{2}(7-1)}(23/7) = -(2/7) = -(-1)^{(7^2-1)/8} = -1. \end{aligned}$$

Or we could have computed $(60/23) = (2^2 \cdot 3 \cdot 5/23) = (3/23)(5/23)$, and then found $(3/23) = (-1)^{11 \cdot 1}(23/3) = -(2/3) = +1$, and $(5/23) = (-1)^{11 \cdot 2}(23/5) = (3/5) = -1$, so that again we find $(60/23) = -1$.

EXERCISES

- 17-1. Find $(-540/11)$, $(540/11)$, $(311/19)$.
- 17-2. Find two solutions of $x^2 \equiv 58 \pmod{23}$.
- 17-3. Prove Corollary 17.2.
- 17-4.* Suppose p is prime. Prove $(-3/p) = 1$ if $p \equiv 1 \pmod{6}$, $(-3/p) = -1$ if $p \equiv 5 \pmod{6}$.
- 17-5. Prove: If p and q are distinct odd primes, then $(p/q) = (q/p)$ if and only if at least one of p, q is congruent to 1 $\pmod{4}$. Consequently, $(p/q) = -(q/p)$ if and only if $p \equiv q \equiv 3 \pmod{4}$.

17-6. Suppose $f(x) = ax^2 + bx + c$, a odd. Show that x_0 is a solution of

$$(18) \quad f(x) \equiv 0 \pmod{2}$$

if and only if x_0 is a solution of $(a + b)x + c \equiv 0 \pmod{2}$. Hence, prove that if $b \equiv 0 \pmod{2}$, (18) always has a solution, and if $b \equiv 1 \pmod{2}$, (18) has a solution if and only if $c \equiv 0 \pmod{2}$. Also, if (18) has no roots, then $a \equiv b \equiv c \equiv 1 \pmod{2}$.

17-7. Prove that $(-1/p) \equiv p \pmod{4}$ for all odd primes p .

18. Congruences with Composite Moduli

If $f(x)$ is a polynomial with integer coefficients and p is a prime, we will show that the solutions (if any) of $f(x) \equiv 0 \pmod{p^b}$ may be used to find all solutions of $f(x) \equiv 0 \pmod{p^{b+1}}$.

Recall that Taylor's Theorem says if $f(x)$ is a polynomial of degree n , if $D^k f$ denotes the k^{th} derivative of f (with the convention that $D^0 f = f$), then

$$f(x_0 + u) = \sum_{k=0}^n \frac{D^k f(x_0)}{k!} u^k.$$

Also, if

$$f(x) = \sum_{k=0}^n c_k x^k$$

and if the c_k are integers, then

$$\begin{aligned} \frac{D^i f(x_0)}{i!} &= \sum_{k=i}^n \frac{k(k-1) \cdots (k-i+1)}{i!} c_k x_0^{k-i} \\ &= \sum_{k=i}^n \binom{k}{i} c_k x_0^{k-i} \end{aligned}$$

is an integer for each i , $0 \leq i \leq n$, whenever $x_0 \in \mathcal{L}$. Therefore, if $x_0, u \in \mathcal{L}$, then

$$f(x_0 + u) = \sum_{k=0}^n C_k u^k$$

where

$$C_k = \frac{D^k f(x_0)}{k!}$$

is an integer.

Theorem 18.1. *If $f(x)$ is a polynomial with integer coefficients, if p is a prime, and if x_0 is a solution of*

$$f(x) \equiv 0 \pmod{p^\beta},$$

then a necessary and sufficient condition that $x_0 + tp^\beta$ be a solution of

$$(19) \quad f(x) \equiv 0 \pmod{p^{\beta+1}}$$

is that t be a solution of

$$(20) \quad Df(x_0)t \equiv -\frac{f(x_0)}{p^\beta} \pmod{p}.$$

Proof. Suppose

$$f(x) = \sum_{k=0}^n c_k x^k$$

and that $x_0 + tp^\beta$ is a solution of (19). By Taylor's Theorem,

$$f(x_0 + tp^\beta) = f(x_0) + Df(x_0)tp^\beta + \sum_{k=2}^n C_k(tp^\beta)^k \equiv 0 \pmod{p^{\beta+1}}$$

so that

$$f(x_0) + Df(x_0)tp^\beta \equiv 0 \pmod{p^{\beta+1}}.$$

But $p^\beta \mid f(x_0)$ by hypothesis, so

$$\frac{f(x_0)}{p^\beta} + Df(x_0)t \equiv 0 \pmod{p},$$

and the condition is necessary.

Conversely, suppose $f(x_0) \equiv 0 \pmod{p^\beta}$ and t is a solution of (20). But

$$f(x_0 + tp^\beta) \equiv f(x_0) + Df(x_0)tp^\beta \pmod{p^{\beta+1}},$$

so $f(x_0 + tp^\beta) \equiv 0 \pmod{p^{\beta+1}}$ and the condition is sufficient.

Corollary 18.1. *Suppose x_0 is a solution of $f(x) \equiv 0 \pmod{p^\beta}$. Corresponding to x_0 , there are $N = N(x_0)$ solutions $x_0 + tp^\beta$ of (19), where $N = 0$ if $p \nmid Df(x_0)$ and $p \nmid \{f(x_0)/p^\beta\}$; $N = p$ if $p \mid Df(x_0)$ and $p \mid \{f(x_0)/p^\beta\}$; and $N = 1$ if $p \nmid Df(x_0)$.*

The proof is left as an exercise.

Example. Find all solutions of $x^3 + x + 3 \equiv 0 \pmod{5^3}$. Clearly, $x_0 = 1$ is the only solution of $f(x) = x^3 + x + 3 \equiv 0 \pmod{5}$; also, $Df(x_0) = (3x^2 + 1)|_{x=1} = 4$, $f(x_0)/5 = 1$. With N denoting the number of solutions of $4t \equiv -1 \pmod{5}$, we have $N = 1$ ($t = 1$). Now $x_0 + tp = 6$ is the only solution of $f(x) \equiv 0 \pmod{5^2}$. With $x_1 = 6$, we find $Df(x_1) = 109$, $f(x_1)/5^2 = 9$,

and N is the number of solutions of $109t \equiv -9 \pmod{5}$, or $4t \equiv 1 \pmod{5}$. Thus, $N = 1$ ($t = 4$). Hence, $x_1 + tp^2 = 106$ is the unique solution of $f(x) \equiv 0 \pmod{5^3}$.

We now consider the general problem of solving a congruence

$$(21) \quad f(x) \equiv 0 \pmod{m}$$

where

$$m = \prod_{i=1}^r p_i^{\beta_i}.$$

Evidently, if x_0 is a solution of (21), then x_0 is a solution of the system

$$(22) \quad \begin{cases} f(x) \equiv 0 \pmod{p_1^{\beta_1}} \\ \dots \\ f(x) \equiv 0 \pmod{p_r^{\beta_r}}. \end{cases}$$

On the other hand, if x_i is a solution of $f(x) \equiv 0 \pmod{p_i^{\beta_i}}$, $i = 1, \dots, r$, then by the Chinese Remainder Theorem there is a solution x_0 of the system

$$\begin{cases} x \equiv x_1 \pmod{p_1^{\beta_1}} \\ \dots \\ x \equiv x_r \pmod{p_r^{\beta_r}}; \end{cases}$$

then x_0 , which is unique $\pmod{m}$, is a solution of (22) and of (21). Therefore, all the solutions of (21) may be obtained by knowing the solutions of $f(x) \equiv 0 \pmod{p_i^{\beta_i}}$ for each i , and we have seen that the solutions of these congruences may be constructed from the solutions of $f(x) \equiv 0 \pmod{p_i}$.

Example. Find all solutions of $f(x) = x^3 + 4x^2 + 3 \equiv 0 \pmod{78408}$. Notice that $78408 = 2^3 3^4 11^2$, and $Df(x) = 3x^2 + 8x$. Let $G_\beta(x, t) = f(x)/p^\beta + Df(x)t$.

We first find solutions of $f(x) \equiv 0 \pmod{2^3}$. Clearly, $a_0 = 1$ is the only solution of $f(x) \equiv 0 \pmod{2}$, and $G_1(a_0, t) = 4 + 11t \equiv 0 \pmod{2}$ has the single solution $t = 0$. Therefore, $a_0 + tp = a_1 = 1$ is the only solution of $f(x) \equiv 0 \pmod{2^2}$. $G_2(a_1, t) = 2 + 11t \equiv 0 \pmod{2}$ has one solution, $t = 0$, so $a_1 + tp^2 = a_2 = 1$ is the only solution of $f(x) \equiv 0 \pmod{2^3}$.

Now, $f(x) \equiv 0 \pmod{3}$ has solutions $b_0 = 0$ and $c_0 = 2$. $G_1(b_0, t) = 1 + 0 \cdot t \equiv 0 \pmod{3}$ has no solution, so there are no solutions of $f(x) \equiv 0 \pmod{3^4}$ obtainable from b_0 . But $G_1(c_0, t) = 9 + 28t \equiv 0 \pmod{3}$ has just one solution, $t = 0$. Therefore, $c_0 + tp = c_1 = 2$ is the only solution of $f(x) \equiv 0 \pmod{3^2}$. $G_2(c_1, t) = 3 + 28t \equiv 0 \pmod{3}$ has $t = 0$ for a solution, and $c_1 + tp^2 = c_2 = 2$ is the only solution of $f(x) \equiv 0 \pmod{3^3}$. $G_3(c_2, t) = 1 + 28t \equiv 0 \pmod{3}$ has $t = 2$ for its solutions, and $c_2 + tp^3 = c_3 = 56$ is the only solution of $f(x) \equiv 0 \pmod{3^4}$.

Finally, $f(x) \equiv 0 \pmod{11}$ has solutions $d_0 = 3, e_0 = 6, g_0 = 9$. $G_1(d_0, t) = 6 + 51t \equiv 0 \pmod{11}$ has $t = 7$ for a solution, and this gives $d_1 = 80$ as a solution of $f(x) \equiv 0 \pmod{11^2}$. $G_1(e_0, t) = 33 + 156t \equiv 0 \pmod{11}$ has $t = 0$ as a solution, and $e_1 = 6$. $G_1(g_0, t) = 96 + 315t \equiv 0 \pmod{11}$ has $t = 2$ for its solution, so $g_1 = 31$.

We now need a solution of $x \equiv a_2 \pmod{2^3}$, $x \equiv c_3 \pmod{3^4}$, $x \equiv d_1 \pmod{11^2}$. By the Chinese Remainder Theorem, $x = 3^4 11^2 + 2^3 11^2 (20)(56) + 2^3 3^4 (76)(80) = 5,033,801 \equiv 15689 \pmod{78408}$ is the solution, and hence is a solution of $f(x) \equiv 0 \pmod{78408}$. Similarly, $y \equiv 1 \pmod{2^3}$, $y \equiv 56 \pmod{3^4}$, $y \equiv 6 \pmod{11^2}$ is solved by $y = 3^4 11^2 + 2^3 11^2 (20)(56) + 2^3 3^4 (76)(6) = 1,389,449 \equiv -21895$. And $z \equiv 1 \pmod{8}$, $z \equiv 56 \pmod{81}$, $z \equiv 31 \pmod{121}$ has the solution $z = 3^4 11^2 + 2^3 11^2 (20)(56) + 2^3 3^4 (76)(31) = 2,620,649 \equiv 33185$.

Therefore, the three solutions of $f(x) \equiv 0 \pmod{78408}$ are 15689, -21895 , and 33185.

EXERCISES

- 18-1. Prove Corollary 18.1. [*Hint*: Use Corollary 12.2 to count the number of solutions of (20).]
- 18-2. Find all solutions of $x^3 + 2x - 3 \equiv 0 \pmod{125}$.
- 18-3. Find all solutions of $x^4 + x^3 + x + 1 \equiv 0 \pmod{2^2 3^3 7^2}$.
- 18-4. Show that if n is even, then $x^2 + x + 1 \equiv 0 \pmod{n}$ has no solution.
- 18-5. Show that if $5|n$, then $f(x) = x^6 + 3 \equiv 0 \pmod{n}$ has no solution. [*Hint*: Consider $f(x) \equiv 0 \pmod{5}$ and use Fermat's Theorem.]

Chapter 4

SUMMATORY FUNCTIONS

19. Introduction

The behavior of many of the arithmetic functions we have studied is quite erratic. For example, if p is prime, then $\tau(p) = 2$, and $\tau(p^k) = k + 1 \rightarrow \infty$ as $n = p^k \rightarrow \infty$. Since there are infinitely many primes, there are infinitely many $n \in \mathcal{Z}^+$ such that $\tau(n) = 2$; on the other hand, for every positive N there are infinitely many n such that $\tau(n) > N$. Thus, there is very little which can be said at this point about the behavior of $\tau(n)$ as n increases without bound. The situation is somewhat more encouraging if we consider the averaging function

$$T_1(n) = \frac{1}{n} \{ \tau(1) + \cdots + \tau(n) \},$$

since averaging the values of the τ function will cause the function T_1 to change less drastically than does the τ function. But studying the function T_1 is equivalent to studying the function

$$T(n) = n T_1(n) = \sum_{d=1}^n \tau(d).$$

Functions of the type $T(n)$ are called *summatory functions*, and in this chapter we study some of the summatory functions associated with the arithmetic functions we have seen previously. A few preliminaries are discussed first. Since $[x]$ is the largest integer not exceeding x , to every real x there corresponds a real number θ_x such that

$$x = [x] + \theta_x, \quad 0 \leq \theta_x < 1.$$

If $f \in \mathcal{A}$, we use the notation

$$\sum_{n \leq x} f(n)$$

to mean

$$\sum_{n=1}^{[x]} f(n);$$

if $x < 1$, this null summation is defined to be zero.

Definition. Let g be a function of a real or integral variable such that $g(x)$ is defined and is positive for all x sufficiently large. The sets $\mathfrak{O}(g)$ and $\mathfrak{o}(g)$ are defined by

$$\mathfrak{O}(g) = \left\{ f: \text{there exists a constant } M \text{ such that} \right. \\ \left. \frac{|f(x)|}{g(x)} < M \text{ for } x \text{ sufficiently large} \right\}, \\ \mathfrak{o}(g) = \left\{ f: \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 0 \right\}.$$

The elements of $\mathfrak{O}(g)$ are denoted $O(g)$ and the elements of $\mathfrak{o}(g)$ are denoted $o(g)$. We refer to both the class $\mathfrak{O}(g)$ and the elements $O(g)$ as “big- O g ”, and $\mathfrak{o}(g)$, $o(g)$ are both read “little- o g ”.

We write $f(x) \sim g(x)$ if and only if

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1;$$

the relation “ $f(x) \sim g(x)$ ” is read “ f is asymptotically equivalent to g .”

For example, $\sin x = O(1)$; that is, $\sin x$ is an element of the class $\mathfrak{O}(1)$, since we can find a constant M , say $M = 2$, such that

$$\frac{|\sin x|}{1} < 2 \quad \text{for all } x.$$

As further examples, we have $x^2 = o(e^x)$, $\sin x = O(x)$, $\sin x = o(x)$, and $x^2 + 2x \sim x^2$.

Since both big- O and little- o terms will arise in equations we will be dealing with, we must assign some convention for combining such terms. Therefore we agree to write

$$O(f) = O(g)$$

if and only if $\mathfrak{O}(f) \subset \mathfrak{O}(g)$. Notice that this is a non-symmetric use of the equality sign. We have, for example, $O(x) = O(x^2)$, but not $O(x^2) = O(x)$. We also agree to denote by $O(f) \pm O(g)$ any element of the class $\mathfrak{O}(h)$ where

$$\mathfrak{O}(h) = \{u + v: u = O(f), v = O(g)\};$$

it should be clear that here $u + v$ is the *sum of functions* defined at each x in the common domain of u and v by $(u + v)(x) = u(x) + v(x)$. We define $O(f) O(g)$ to be any element of the class

$$\mathfrak{O}(h) = \{uv : u = O(f), v = O(g)\},$$

and uv is defined by $uv(x) = u(x)v(x)$. Similar rules hold for combining little- o terms.

As an immediate consequence of these definitions, we have $O(f) = O(g)$ if and only if $h = O(f)$ implies $h = O(g)$. Also, $O(f) = O(g)$ if and only if $f = O(g)$.

The following properties hold.

- (1) $O(g) \pm O(g) = O(g)$
- (2) $o(g) \pm o(g) = o(g)$
- (3) $O(O(g)) = O(g)$
- (4) $O(o(g)) = o(g)$
- (5) $\{O(g)\}^2 = O(g^2)$.

To prove (1), suppose $f = O(g)$ and $h = O(g)$. Then there are constants M_1 and M_2 such that $|f(x)| < M_1 g(x)$ and $|h(x)| < M_2 g(x)$, from which we have

$$|f(x) \pm h(x)| \leq |f(x)| + |h(x)| < (M_1 + M_2)g(x),$$

so that $f \pm h = O(g)$.

Property (4) may be proved as follows. Let $f = o(g)$ and $h = O(f)$, so that $h = O(o(g))$. There is a constant M such that $|h(x)| < Mf(x)$ and $f(x)/g(x) \rightarrow 0$. Then

$$0 \leq \lim_{x \rightarrow \infty} \frac{|h(x)|}{g(x)} \leq M \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 0,$$

and $h = o(g)$.

It need not be true in general that

$$\sum_{i=1}^n O(g_i(x)) = O\left(\sum_{i=1}^n g_i(x)\right).$$

For example, for the sequence of constant functions $f_i(u) = i$, we have $f_1(u) = O(1)$, $f_2(u) = O(1)$, $f_3(u) = O(1)$, ..., but

$$\sum_{i \leq x} f_i(u) = \frac{[x]([x] + 1)}{2} = O(x^2) \neq O\left(\sum_{i \leq x} 1\right) = O(x).$$

A sufficient condition is that the constants implied by the O symbols be uniformly bounded. With such a restriction we may add infinitely many error terms $O(g_i(x))$.

Suppose $|f_i(x)| < M_i g_i(x)$ for $i = 1, 2, \dots$, that there exists a constant M such that $M_i < M$ for all i , and that

$$\sum_{i=1}^{\infty} g_i(x)$$

converges for every $x \geq x_0$. Then

$$(6) \quad \sum_{i=1}^{\infty} O(g_i(x)) = O\left(\sum_{i=1}^{\infty} g_i(x)\right).$$

This is because $|\sum f_i(x)| \leq \sum |f_i(x)| \leq M \sum g_i(x)$. Obviously, (6) also holds if the series are replaced by finite sums.

The above list of properties is certainly not exhaustive, but contains typical rules for combining terms involving the symbols O and o . It is usually much easier to derive whatever rules are necessary in some particular application than to memorize a lengthy list of such rules.

Lemma 19.1. *There is a constant γ , called Euler's constant, such that $1/2 < \gamma < 1$ and*

$$\sum_{n \leq x} \frac{1}{n} = \log x + \gamma + O\left(\frac{1}{x}\right).$$

Proof. We first prove the lemma for integers x , and then extend the result to the general case. Thus, assume $x = N$ is an integer. We define

$$\gamma_N = \sum_{n=1}^N \frac{1}{n} - \log N, \quad N = 1, 2, 3, \dots$$

We note that the sequence $\{\gamma_N\}$ is convergent, because

$$\begin{aligned} \gamma_N - \gamma_{N+1} &= \log(N+1) - \log(N) - \frac{1}{N+1} \\ &= \int_N^{N+1} \frac{dx}{x} - \frac{1}{N+1}. \end{aligned}$$

Geometrically, this is the area of the plane region bounded by the curve $y = 1/x$, the x -axis, and the lines $x = N$ and $x = N+1$ minus the area of a rectangle of length 1 and height $1/(N+1)$. Since a rectangle of this type may be inscribed in the region with area $\int_N^{N+1} x^{-1} dx$, the difference is positive (the shaded area in Figure 2). Hence, $\gamma_N - \gamma_{N+1} > 0$, so the sequence is decreasing.

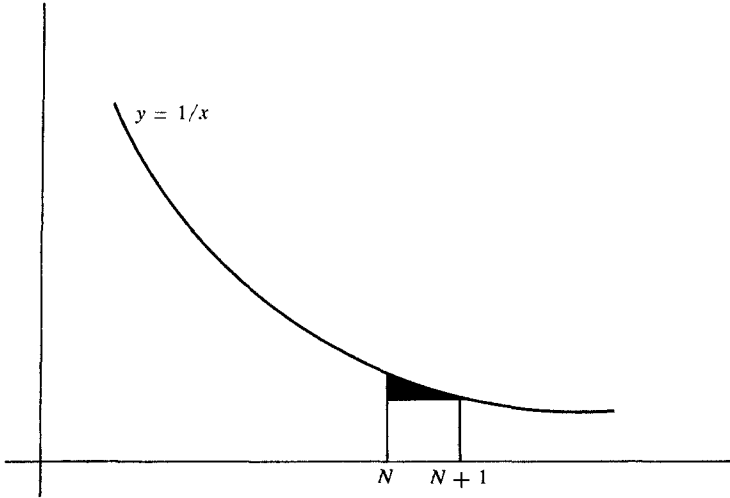


Figure 2

Furthermore, for every N , $\gamma_N > 0$, which can be seen as follows. We have

$$\begin{aligned}
 \gamma_N &= 1 - \left\{ \log N - \sum_{n=2}^N \frac{1}{n} \right\} \\
 &= 1 - \left\{ \int_1^N \frac{dx}{x} - \sum_{n=2}^N \frac{1}{n} \right\} \\
 &= 1 - \left\{ \sum_{n=2}^N \int_{n-1}^n \frac{dx}{x} - \sum_{n=2}^N \frac{1}{n} \right\} \\
 &= 1 - \sum_{n=2}^N \left\{ \int_{n-1}^n \frac{dx}{x} - \frac{1}{n} \right\}.
 \end{aligned}$$

Geometrically, this is the area of a unit square *minus* the sum of shaded areas of the type in Figure 2 for $n = 2, \dots, N$. These shaded areas may be translated to the left by $n - 1$ units; since $y = 1/x$ is continuously decreasing, these areas $\int_{n-1}^n x^{-1} dx - 1/n$ will not overlap when translated into the unit square $0 \leq y \leq 1, 0 \leq x \leq 1$ and then γ_N is the (positive) area of the unshaded portion of the unit square. The argument is illustrated for γ_4 in Figure 3.

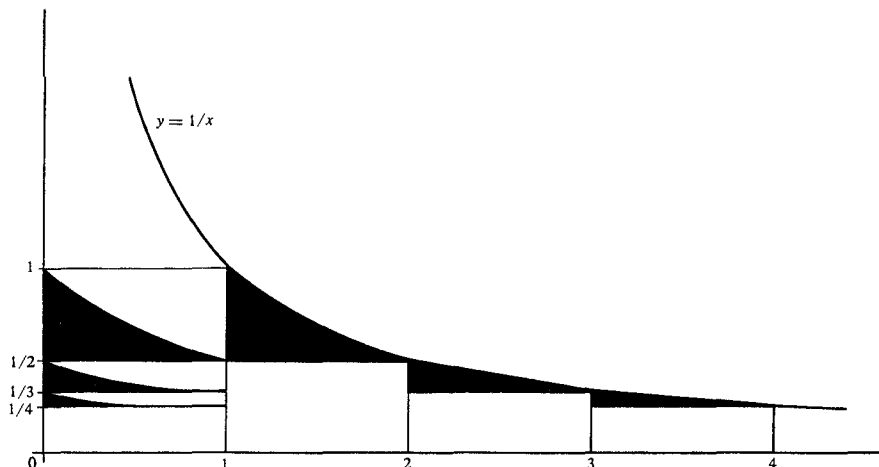


Figure 3

Since $\gamma_N > 0$ for all N and the sequence $\{\gamma_N\}$ is decreasing, it is convergent, say

$$\lim_{n \rightarrow \infty} \gamma_n = \gamma.$$

From our arguments above it is clear that γ is the limit of the unshaded region in the unit square under the translations described. Since this unshaded region is in the square, we evidently have $\gamma < 1$. Also, it is geometrically obvious that γ is the sum of a series whose N^{th} term a_N is the area of the unshaded portion of the rectangle $0 \leq x \leq 1$, $1/N \geq y \geq 1/(N+1)$. Since this unshaded portion of the rectangle exceeds half the area of the rectangle (the shaded part of each rectangle lies entirely below the diagonal), we have

$$a_N > \frac{1}{2} \left(\frac{1}{N} - \frac{1}{N+1} \right) = \frac{1}{2N(N+1)}.$$

Therefore,

$$\gamma = \sum_{N=1}^{\infty} a_N > \frac{1}{2} \sum_{N=1}^{\infty} \frac{1}{N(N+1)} = \frac{1}{2}.$$

Finally, we consider $\gamma_N - \gamma$. By the translation of shaded areas we know this is the area of the unshaded portion of the unit square inside the region $0 \leq x \leq 1$, $1/N \leq y \leq 1$, minus the limit of the unshaded areas. But this difference lies within the region $0 \leq x \leq 1$, $0 \leq y \leq 1/N$ with area $1/N$.

Hence $\gamma_N - \gamma < 1/N$, or $\gamma_N - \gamma = O(1/N)$; applying definitions, we have

$$\sum_{n=1}^N \frac{1}{n} = \log N + \gamma + O\left(\frac{1}{N}\right),$$

which is the lemma in case x is an integer.

Now suppose $x \geq 1$ is arbitrary (not necessarily an integer). By what we have already proved,

$$(7) \quad \sum_{n \leq x} \frac{1}{n} = \sum_{n=1}^{[x]} \frac{1}{n} = \log [x] + \gamma + O\left(\frac{1}{[x]}\right).$$

Since $[x] = x + O(1)$, we have

$$\log [x] = \log (x + O(1)) = \log \left(x \left\{ 1 + O\left(\frac{1}{x}\right) \right\} \right)$$

$$(8) \quad \log [x] = \log x + \log \left(1 + O\left(\frac{1}{x}\right) \right).$$

Now recall that for $|t| < 1$,

$$\log (1 + t) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} t^n}{n}$$

so that

$$\begin{aligned} \log \left(1 + O\left(\frac{1}{x}\right) \right) &= \sum_{n=1}^{\infty} \frac{(-1)^{n-1} (O(1/x))^n}{n} \\ &= \sum_{n=1}^{\infty} O(1/x^n) \\ &= O\left(\sum_{n=1}^{\infty} \frac{1}{x^n} \right) \end{aligned}$$

since the constant involved in $O(1/x^n)$ is θ_x^n/n , and these are uniformly less than 1. Therefore,

$$\log \left(1 + O\left(\frac{1}{x}\right) \right) = O\left(\frac{1}{x} \frac{x}{x-1} \right) = O\left(\frac{1}{x}\right).$$

Thus, from (8) we have $\log [x] = \log x + O(1/x)$. These results together with Exercise 19-2 into (7) give us

$$\begin{aligned} \sum_{n \leq x} \frac{1}{n} &= \log x + O(1/x) + \gamma + O(1/x) \\ &= \log x + \gamma + O(1/x). \end{aligned}$$

The proof is complete.

Note. Lemma 19.1 also implies (in a trivial way) certain weaker asymptotic formulas. For example,

$$\sum_{n \leq x} \frac{1}{n} = \log x + O(1),$$

because $\gamma = O(1)$ and $O(1/x) = O(1)$. In many applications, we lose nothing by using a weaker result of this type. For example, see the Note after Theorem 21.1.

EXERCISES

- 19-1. Prove properties (2), (3), and (5).
- 19-2. Show that $O(1/\lfloor x \rfloor) = O(1/x)$.
- 19-3. Prove that “ $\sim$ ” is an equivalence relation on the set of positive-valued functions.
- 19-4. Prove the big- O relation is reflexive and transitive, but not symmetric, on the set of positive-valued functions.
- 19-5. Prove that the little- o relation is transitive. Show by examples that it is neither reflexive nor symmetric.
- 19-6.* Prove

$$\sum_{n \leq x} \frac{1}{n} = O(\log x).$$

- 19-7. Prove the remarks made earlier, namely that the following conditions are equivalent.
 - (a) $O(f) = O(g)$;
 - (b) $h = O(f)$ implies $h = O(g)$;
 - (c) $f = O(g)$.
- 19-8. Prove the following conditions are equivalent:
 - (a) $o(f) = o(g)$;
 - (b) if $h = o(f)$, then $h = o(g)$;
- 19-9. Prove that $O(O(O(g))) = O(g)$.
- 19-10. Prove $o(o(g)) = o(g)$.

20. The Euler–McLaurin Sum Formula

It is interesting to note that techniques similar to those employed in the proof of Lemma 19.1 may be applied in a more general setting to obtain a convenient tool for dealing with sums of the type under consideration in this chapter. While we will not need this result until later, we now prove Theorem 20.1.

Theorem 20.1. (*Euler–McLaurin sum formula.*) Suppose $f(x)$ is continuously differentiable for $x \geq 1$. Then

$$\sum_{n \leq x} f(n) = f(1) + \int_1^x f(t) dt + \int_1^x (t - [t]) Df(t) dt - (x - [x]) f(x).$$

Note. In many applications it is sufficient to replace the last term by $O(f)$.

Proof. We first prove the theorem when $x = N$ is an integer. Then

$$\begin{aligned} \sum_{n \leq N} f(n) &= N f(N) - \sum_{n=2}^N (n-1) \{f(n) - f(n-1)\} \\ &= N f(N) - \sum_{n=2}^N (n-1) \int_{n-1}^n Df(t) dt \\ &= N f(N) - \sum_{n=2}^N \int_{n-1}^n [t] Df(t) dt \end{aligned}$$

$$(9) \quad \sum_{n \leq N} f(n) = N f(N) - \int_1^N [t] Df(t) dt.$$

But, using integration by parts,

$$\int_1^N f(t) dt = t f(t) \Big|_1^N - \int_1^N t Df(t) dt = N f(N) - f(1) - \int_1^N t Df(t) dt.$$

Substituting into (9) the expression for $N f(N)$ from this equation, we have

$$\sum_{n \leq N} f(n) = f(1) + \int_1^N f(t) dt + \int_1^N (t - [t]) Df(t) dt,$$

which is the result in Theorem 20.1 when $x = [x]$ is an integer.

Now suppose x is arbitrary. Then

$$\begin{aligned} \sum_{n \leq x} f(n) &= f(1) + \int_1^{[x]} f(t) dt + \int_1^{[x]} (t - [t]) Df(t) dt \\ &= f(1) + \int_1^x f(t) dt + \int_1^x (t - [t]) Df(t) dt \\ &\quad - \left\{ \int_{[x]}^x f(t) dt + \int_{[x]}^x (t - [t]) Df(t) dt \right\}. \end{aligned}$$

But

$$\begin{aligned} \int_{[x]}^x f(t) dt + \int_{[x]}^x t Df(t) dt - \int_{[x]}^x [t] Df(t) dt \\ = t f(t) \Big|_{[x]}^x - [x] f(t) \Big|_{[x]}^x = (x - [x]) f(x); \end{aligned}$$

here we have used integration by parts and observed that for $[x] \leq t \leq x$, $[t] = [x]$ is constant. The proof is complete.

Example. Use Theorem 20.1 to prove there is a constant C_1 such that

$$\sum_{n \leq x} \frac{1}{n} = \log x + C_1 + O(1/x).$$

In Theorem 20.1, take $f(t) = 1/t$. Then $Df(t) = -1/t^2$ and

$$\sum_{n \leq x} \frac{1}{n} = 1 + \int_1^x \frac{dt}{t} - \int_1^x \frac{(t - [t])}{t^2} dt - \frac{x - [x]}{x}.$$

Clearly

$$\frac{x - [x]}{x} = O(1/x).$$

Also, if the following improper integrals are convergent, we may write

$$(10) \quad \sum_{n \leq x} \frac{1}{n} = 1 + \log x - \int_1^\infty \frac{t - [t]}{t^2} dt + \int_x^\infty \frac{t - [t]}{t^2} dt + O\left(\frac{1}{x}\right).$$

But the first integral is convergent if the second approaches 0 as $x \rightarrow \infty$, and we see that this is the case since

$$0 \leq \int_x^\infty \frac{t - [t]}{t^2} dt \leq \int_x^\infty \frac{dt}{t^2} = \frac{1}{x}.$$

Therefore

$$\int_x^\infty \frac{t - [t]}{t^2} dt = O\left(\frac{1}{x}\right)$$

and

$$\int_1^\infty \frac{t - [t]}{t^2} dt$$

converges, say with value $1 - C_1$. Substitution into (10) proves the result.

EXERCISES

20-1. We have defined

$$C_1 = 1 - \int_1^\infty \frac{t - [t]}{t^2} dt$$

$$\gamma = \lim_{N \rightarrow \infty} \left\{ \sum_{n=1}^N \frac{1}{n} - \log N \right\}.$$

Prove that $C_1 = \gamma$.

20-2. Suppose N is an integer, $1 \leq N \leq x$, and $f(t)$ is continuously differentiable for $t \geq 1$. Prove that

$$\sum_{n=N}^x f(n) = f(N) + \int_N^x f(t) dt + \int_N^x (t - [t]) Df(t) dt + O(f).$$

20-3.* Prove that

$$\sum_{n \leq x} \frac{1}{n^2} = K - \frac{1}{x} + O(1/x^2)$$

for some constant K , $1 < K < 2$. It is known that $K = \pi^2/6$.

20-4. If $k \in \mathcal{Z}^+$, $k > 1$, prove

$$\sum_{n \leq x} \frac{1}{n^k} = \frac{k}{k-1} + O(x^{-k+1}).$$

20-5. Notice that $0 \leq t - [t] < 1$ for all real t . Use this to show that

$$\int_1^x (t - [t]) t^{k-1} dt = O(x^k)$$

for all $k \in \mathcal{Z}^+$. Hence use the Euler-McLaurin sum formula to prove that for every $k \in \mathcal{Z}^+$

$$\sum_{n \leq x} n^k = C(k) x^{k+1} + O(x^k)$$

where $C(k)$ is a constant depending only on k , not on x .

21. Order of Magnitude of $\tau(n)$

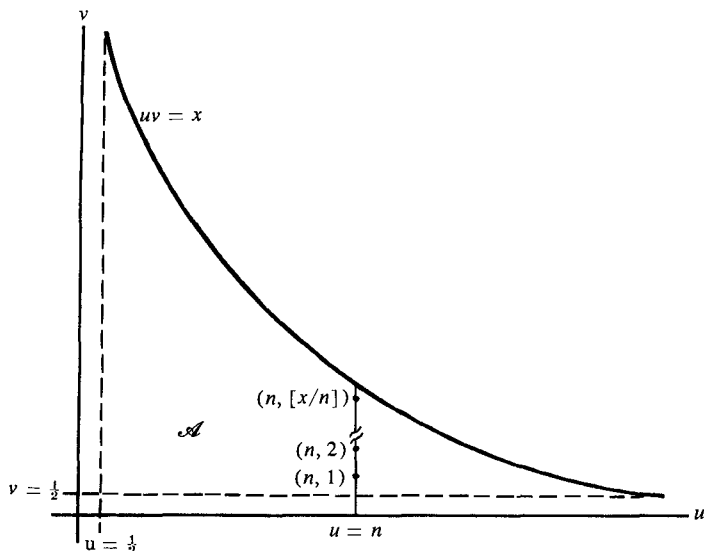
Theorem 21.1.

$$\sum_{n \leq x} \tau(n) = x \log x + O(x)$$

Proof. Let the plane region $\mathcal{A}$ be the region bounded by the lines $u = 1/2$, $v = 1/2$, and the hyperbola $uv = x$ (see Figure 4), where x is arbitrary but fixed. We will show that

$$\sum_{n \leq x} \tau(n) = \Lambda_{\mathcal{A}} = \sum_{n \leq x} [x/n].$$

If $1 \leq n \leq x$, then $\tau(n)$ is just the number of lattice points on the hyperbola $uv = n$, because corresponding to each positive divisor d of n , there is an integer n/d such that $(u, v) = (d, n/d)$ is a lattice point on $uv = n$. Thus, the sum $\sum_{n \leq x} \tau(n)$ represents the number of lattice points in the first quadrant on any hyperbola $uv = n$, $1 \leq n \leq x$. But every lattice point in $\mathcal{A}$ is on exactly one such hyperbola, so $\sum_{n \leq x} \tau(n) = \Lambda_{\mathcal{A}}$.

**Figure 4**

Again with $1 \leq n \leq x$, consider the number of lattice points in $\mathcal{A}$ and on the line $u = n$ (see Figure 4). These are just the points $(n, 1), \dots, (n, \xi)$, where ξ is the largest integer such that $n\xi \leq x$, or $\xi = [x/n]$. Since all lattice points in $\mathcal{A}$ are on exactly one such line, $\Lambda_{\mathcal{A}} = \sum_{n \leq x} [x/n]$. Therefore,

$$\begin{aligned}
 \sum_{n \leq x} \tau(n) &= \sum_{n \leq x} [x/n] \\
 &= \sum_{n \leq x} \left\{ \frac{x}{n} + O(1) \right\} \\
 &= x \sum_{n \leq x} \frac{1}{n} + O\left(\sum_{n \leq x} 1 \right) = x \left\{ \log x + \gamma + O\left(\frac{1}{x} \right) \right\} + O(x) \\
 &= x \log x + \gamma x + O(1) + O(x) = x \log x + O(x).
 \end{aligned}$$

Note. In the above, we could have used the weaker result given in the Note at the end of Section 19 and we would have obtained the same result. However, the still weaker form given in Exercise 19-6 could *not* have been used to obtain this same theorem.

By using a refinement of the techniques employed above we can prove a much sharper result.

Theorem 21.2.

$$\sum_{n \leq x} \tau(n) = x(\log x + 2\gamma - 1) + O(\sqrt{x})$$

Proof. We have already shown that $\sum_{n \leq x} \tau(n) = \Lambda_{\mathcal{A}}$ where $\mathcal{A}$ is the plane region described in the proof of Theorem 21.1. Now we divide $\mathcal{A}$ into sub-regions $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ as follows: $\mathcal{B}$ is bounded by $v = 1/2$, $u = 1/2$, $u = [\sqrt{x}]$, $uv = x$; $\mathcal{C}$ is bounded by $u = 1/2$, $v = 1/2$, $v = [\sqrt{x}]$, $uv = x$; $\mathcal{D}$ is bounded by $u = [\sqrt{x}]$, $v = [\sqrt{x}]$, $uv = x$. See Figure 5.

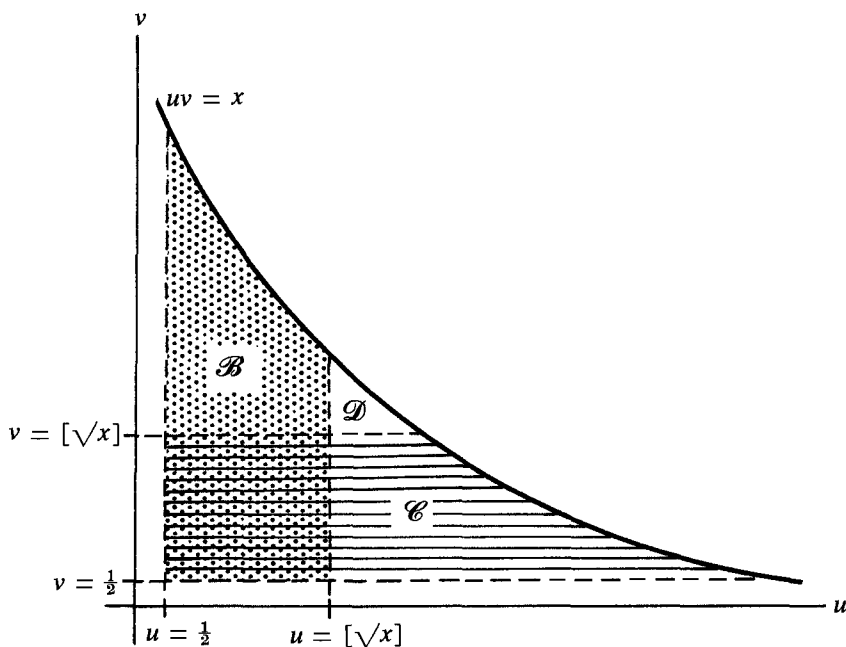


Figure 5

Evidently, $\Lambda_{\mathcal{A}} = \Lambda_{\mathcal{B}} + \Lambda_{\mathcal{C}} - \Lambda_{\mathcal{B} \cap \mathcal{C}} + \Lambda_{\mathcal{D}} = 2\Lambda_{\mathcal{B}} - \Lambda_{\mathcal{B} \cap \mathcal{C}} + \Lambda_{\mathcal{D}}$. Since $\mathcal{B} \cap \mathcal{C}$ is a square it is clear that $\Lambda_{\mathcal{B} \cap \mathcal{C}} = [\sqrt{x}]^2 = (\sqrt{x} + O(1))^2 = x + O(\sqrt{x})$. Also, $\mathcal{D} \cap \mathcal{C}$ is the line segment from $([\sqrt{x}], [\sqrt{x}])$ to $(x/[\sqrt{x}], [\sqrt{x}])$, with length $x/[\sqrt{x}] - [\sqrt{x}] < 3$; thus $\mathcal{D}$ may be inscribed in a square with sides of length 3, and $\Lambda_{\mathcal{D}}$ cannot exceed the area of such a square, so $\Lambda_{\mathcal{D}} = O(1)$.

Finally,

$$\begin{aligned}
 2\Lambda_{\mathscr{B}} &= 2 \sum_{n \leq \sqrt{x}} [x/n] = 2 \sum_{n \leq \sqrt{x}} \left\{ \frac{x}{n} + O(1) \right\} \\
 &= 2x \sum_{n \leq \sqrt{x}} \frac{1}{n} + O\left(\sum_{n \leq \sqrt{x}} 1 \right) \\
 &= 2x(\log \sqrt{x} + \gamma + O(1/\sqrt{x})) + O(\sqrt{x}) \\
 &= x \log x + 2\gamma x + O(\sqrt{x}).
 \end{aligned}$$

Combining results we have

$$\begin{aligned}
 \sum_{n \leq x} \tau(n) &= x \log x + 2\gamma x + O(\sqrt{x}) - x - O(\sqrt{x}) + O(1) \\
 &= x(\log x + 2\gamma - 1) + O(\sqrt{x}).
 \end{aligned}$$

The results in either of the last two theorems may be stated loosely by saying that $\tau(n)$ has average order of magnitude equal to $\log n$; that is,

$$\frac{1}{N} \sum_{n \leq N} \tau(n) \sim \log N.$$

This relation follows easily from either theorem: since

$$T(N) = \sum_{n \leq N} \tau(n) = N \log N + O(N),$$

we have $(1/N)T(N) = \log N + O(1)$, and

$$\lim_{N \rightarrow \infty} \frac{(1/N)T(N)}{\log N} = \lim_{N \rightarrow \infty} \frac{\log N + O(1)}{\log N} = 1.$$

22. Order of Magnitude of $\sigma(n)$

In the proofs of the last two theorems the same geometric argument was used to evaluate a function at the lattice points in a plane region in two different ways. The technique has a general applicability and may be used to prove the following.

Theorem 22.1.

$$\sum_{n \leq x} f \cdot g(n) = \sum_{d \leq x} f(d) \sum_{n \leq x/d} g(n)$$

Proof.

$$\sum_{n \leq x} f \cdot g(n) = \sum_{n \leq x} \sum_{ab=n} f(a) g(b)$$

is the sum of the evaluations of $f(a)g(b)$ at each lattice point on the upper branch of $uv = n$ for every n , $1 \leq n \leq x$, so the evaluation extends over every lattice point in the region $\mathcal{A}$ of Figure 4.

Now let $d \in \mathcal{L}$ be fixed, $1 \leq d \leq x$, and consider the sum of the evaluations of $f(d)g(n)$ at every lattice point in $\mathcal{A}$ which is on the line $u = d$. This is just $\sum f(d)g(n)$ summed over those n such that $dn \leq x$, or equivalently,

$$f(d) \sum_{n \leq x/d} g(n).$$

Summing this last expression over d , $1 \leq d \leq x$, proves the theorem.

Example. We have already seen the special case of Theorem 22.1 obtained by taking $f = g = \iota_0$, namely

$$\begin{aligned} \sum_{n \leq x} \tau(n) &= \sum_{n \leq x} \iota_0 \cdot \iota_0(n) = \sum_{d \leq x} \iota_0(d) \sum_{n \leq x/d} \iota_0(n) \\ &= \sum_{d \leq x} 1 \sum_{n \leq x/d} 1 = \sum_{d \leq x} [x/d]. \end{aligned}$$

Example. In Theorem 22.1, take $f = \iota_1$ and $g = \iota_0$. We get

$$\begin{aligned} \sum_{n \leq x} \sigma(n) &= \sum_{d \leq x} \iota_1(d) \sum_{n \leq x/d} \iota_0(n) = \sum_{d \leq x} d[x/d] \\ &= \sum_{d \leq x} d(x/d + O(1)) = \sum_{d \leq x} x + O\left(\sum_{d \leq x} d\right) \\ &= x[x] + O(x^2) = O(x^2). \end{aligned}$$

In contrast to the last example, we now show that a sharper result is obtainable by using the commutativity of the product of arithmetic functions.

Theorem 22.2.

$$\sum_{n \leq x} \sigma(n) = C_2 x^2 + O(x \log x)$$

for some constant C_2 , $1/2 < C_2 < 1$.

Proof. Since $\sigma = \iota_0 \cdot \iota_1$, we have by Theorem 22.1

$$\begin{aligned} \sum_{n \leq x} \sigma(n) &= \sum_{d \leq x} 1 \sum_{n \leq x/d} n \\ &= \sum_{d \leq x} \frac{[x/d]([x/d] + 1)}{2} \\ &= \frac{1}{2} \sum_{d \leq x} \left(\frac{x}{d} + O(1) \right)^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{x^2}{2} \sum_{d \leq x} \frac{1}{d^2} + O\left(x \sum_{d \leq x} \frac{1}{d}\right) + O\left(\sum_{d \leq x} 1\right) \\
&= \frac{x^2}{2} \left(K - \frac{1}{x} + O(1/x^2)\right) + O(x \log x), \quad \text{from Exercise 20-3,} \\
&= \frac{Kx^2}{2} - \frac{x}{2} + O(1) + O(x \log x).
\end{aligned}$$

Replacing $K/2$ by C_2 and noting that $-x/2 + O(1) = O(x \log x)$, we have the theorem proved.

EXERCISES

22-1. Prove that $x \log x = o(x^2) = O(x^2)$. This shows that the error term $O(x \log x)$ in Theorem 22.2 is of smaller order of magnitude than the dominant term $C_2 x^2$.

22-2. Prove that for $k \in \mathcal{L}^+$,

$$\sum_{n \leq x} \sigma_k(n) = O(x^{k+1}).$$

(Hint: Use Theorem 22.1 and the result in Exercise 20-5.)

23. Sums Involving the Möbius Function

To motivate the introduction of further machinery, consider an application of Theorem 22.1 to find an asymptotic formula for the function $\Sigma \varphi(n)$. Since $\varphi = \iota_1 \cdot \mu = \mu \cdot \iota_1$,

$$\sum_{n \leq x} \varphi(n) = \sum_{n \leq x} \iota_1 \cdot \mu(n) = \sum_{d \leq x} \iota_1(d) \sum_{n \leq x/d} \mu(n)$$

and we cannot proceed without some knowledge of the function $\Sigma_{n \leq x} \mu(n)$. By applying the identity $\varphi = \mu \cdot \iota_1$, we get

$$\begin{aligned}
\sum_{n \leq x} \varphi(n) &= \sum_{d \leq x} \mu(d) \sum_{n \leq x/d} \iota_1(n) = \sum_{d \leq x} \mu(d) \sum_{n \leq x/d} n \\
&= \sum_{d \leq x} \mu(d) \frac{[x/d]([x/d] + 1)}{2} \\
(11) \quad &= \frac{x^2}{2} \sum_{d \leq x} \frac{\mu(d)}{d^2} + O\left(x \sum_{d \leq x} \frac{|\mu(d)|}{d}\right) + O\left(\sum_{d \leq x} |\mu(d)|\right).
\end{aligned}$$

It now appears that we need to know $\Sigma \mu(n) n^{-k}$ for $k = 0, 1$, and 2 . Such sums involving the function μ can be evaluated with the use of our next two theorems.

Theorem 23.1. Suppose f and g are functions of a real variable. If

$$(12) \quad f(x) = \sum_{n \leq x} g\left(\frac{x}{n}\right),$$

then

$$(13) \quad g(x) = \sum_{n \leq x} \mu(n) f\left(\frac{x}{n}\right).$$

Conversely, (13) implies (12).

Proof. Suppose (12) holds. Then

$$\sum_{d \leq x} \mu(d) f\left(\frac{x}{d}\right) = \sum_{d \leq x} \mu(d) \sum_{n \leq x/d} g\left(\frac{x}{nd}\right).$$

As in the proof of Theorem 22.1 we see that this is simply the sum of evaluations of $\mu(d)g(x/nd)$ at each lattice point (n, d) in the region bounded by $u = 1/2$, $v = 1/2$, $uw = x$. Changing the order of summation we evaluate $\mu(a)g(x/ab)$ at each lattice point (a, b) on $uv = n$, $1 \leq n \leq x$, and then sum from $n = 1$ to $n = [x]$. Therefore,

$$\begin{aligned} \sum_{d \leq x} \mu(d) f\left(\frac{x}{d}\right) &= \sum_{n \leq x} \sum_{ab=n} \mu(a) g\left(\frac{x}{ab}\right) \\ &= \sum_{n \leq x} g\left(\frac{x}{n}\right) \sum_{ab=n} \mu(a) \\ &= \sum_{n \leq x} g\left(\frac{x}{n}\right) \varepsilon(n) = g(x) \end{aligned}$$

since $\varepsilon(n) = 1$ if and only if $n = 1$, zero otherwise. This prove (13).

Conversely, assume (13). Then

$$\begin{aligned} \sum_{d \leq x} g\left(\frac{x}{d}\right) &= \sum_{d \leq x} \sum_{n \leq x/d} \mu(n) f\left(\frac{x}{nd}\right) = \sum_{n \leq x} \sum_{ab=n} \mu(a) f\left(\frac{x}{ab}\right) \\ &= \sum_{n \leq x} f\left(\frac{x}{n}\right) \varepsilon(n) = f(x). \end{aligned}$$

Corollary 23.1.

$$\sum_{n \leq x} \frac{\mu(n)}{n} = O(1), \quad \sum_{n \leq x} \frac{|\mu(n)|}{n} = O(\log x).$$

Proof. Take $g(x) = 1$. Then

$$f(x) = \sum_{n \leq x} 1 = x + O(1).$$

By the inversion formula of the theorem, we have

$$\begin{aligned} 1 &= \sum_{n \leq x} \mu(n) f\left(\frac{x}{n}\right) = \sum_{n \leq x} \mu(n) \left\{ \frac{x}{n} + O(1) \right\} \\ &= x \sum_{n \leq x} \frac{\mu(n)}{n} + \sum_{n \leq x} \mu(n) O(1). \end{aligned}$$

But since $|\mu(n)| \leq 1$,

$$\sum_{n \leq x} \mu(n) O(1) = O(x)$$

and

$$1 = x \sum_{n \leq x} \frac{\mu(n)}{n} + O(x)$$

or

$$\sum_{n \leq x} \frac{\mu(n)}{n} = \frac{1}{x} + O(1) = O(1).$$

We now introduce the Dirichlet series, which will be used to study other sums involving the Möbius function. If $f \in \mathcal{A}$, then a series of the type

$$\sum_{n=1}^{\infty} \frac{f(n)}{n^s}, \quad s \text{ real},$$

is called a *Dirichlet series*. The convergence or divergence of these series is determined by the function f and the value of s . It is easy to prove if $f(n) = O(1)$ and $s > 1$, then

$$\sum_{n=1}^{\infty} f(n) n^{-s}$$

is absolutely convergent. If

$$F(s) = \sum_{n=1}^{\infty} f(n) n^{-s} \quad \text{and} \quad G(s) = \sum_{n=1}^{\infty} g(n) n^{-s}$$

are two Dirichlet series, then the product $F(s) G(s)$ is defined by

$$(14) \quad F(s) G(s) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} f(n) g(m) n^{-s} m^{-s}$$

if this series is convergent.

Theorem 23.2. *If $F(s) = \sum f(n) n^{-s}$ and $G(s) = \sum g(n) n^{-s}$ are two Dirichlet series which are absolutely convergent, then $F(s) G(s)$ is defined and*

$$F(s) G(s) = \sum_{n=1}^{\infty} \frac{f \cdot g(n)}{n^s}.$$

Proof. It is known that the product series is absolutely convergent under the given conditions, so the terms in the product series may be rearranged without affecting the sum. Let $n \in \mathcal{Z}^+$ be arbitrary. If $n = ab$ is any factorization of n into positive integers, the series (14) contains a term

$$f(a)g(b)a^{-s}b^{-s} = \frac{f(a)g(b)}{n^s}.$$

Selecting out of (14) all such terms for factorizations of n , we see that (14) contains

$$\sum_{ab=n} \frac{f(a)g(b)}{n^s} = f \cdot g(n)n^{-s}$$

and no other terms with n^{-s} involved. Since n was arbitrary, we have

$$F(s)G(s) = \sum_{n=1}^{\infty} \frac{f \cdot g(n)}{n^s}.$$

Remark. If $s > 1$, the Dirichlet series $\zeta(s) = \sum \iota_0(n)n^{-s} = \sum 1/n^s$ is the Riemann zeta function. It is known that $\zeta(2) = \pi^2/6$ and $\zeta(4) = \pi^4/90$.

Corollary 23.2a.

$$\sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} = \frac{1}{\zeta(s)}, \quad s > 1$$

Proof.

$$\left\{ \sum \frac{\mu(n)}{n^s} \right\} \left\{ \sum \frac{\iota_0(n)}{n^s} \right\} = \sum \frac{\mu \cdot \iota_0(n)}{n^s} = \sum \frac{\varepsilon(n)}{n^s}$$

Applying definitions, we have $(\sum \mu(n)n^{-s})\zeta(s) = 1$.

Corollary 23.2b.

$$\sum_{n \leq x} \frac{\mu(n)}{n^2} = \frac{6}{\pi^2} + O\left(\frac{1}{x}\right)$$

Proof.

$$\begin{aligned} \sum_{n \leq x} \frac{\mu(n)}{n^2} &= \sum_{n=1}^{\infty} \frac{\mu(n)}{n^2} - \sum_{n=\{x\}+1}^{\infty} \frac{\mu(n)}{n^2} \\ &= \frac{1}{\zeta(2)} + O\left(\sum_{n>x} \frac{1}{n^2}\right) \\ &= \frac{6}{\pi^2} + O\left(\int_x^{\infty} \frac{dt}{t^2}\right) \\ &= \frac{6}{\pi^2} + O\left(\frac{1}{x}\right). \end{aligned}$$

Putting the appropriate results into (11) we obtain the following.

Theorem 23.3.

$$\sum_{n \leq x} \varphi(n) = \frac{3x^2}{\pi^2} + O(x \log x)$$

Our next theorem is an example of a somewhat different type of application of Theorem 23.2. Before proceeding, the reader should review Theorem 14.3.

Theorem 23.4. *If $s > 1$ and if $c_n(d)$ is Ramanujan's trigonometric function, then*

$$\frac{\sigma_{s-1}(n)}{n^{s-1}} = \zeta(s) \sum_{d=1}^{\infty} \frac{c_n(d)}{d^s}.$$

Proof.

$$\begin{aligned} \zeta(s) \sum_{d=1}^{\infty} \frac{c_n(d)}{d^s} &= \sum_{d=1}^{\infty} \frac{\iota_0 \cdot c_n(d)}{d^s} = \sum_{d|n} \frac{d}{d^s} \\ &= \frac{1}{n^{s-1}} \sum_{d|n} \left(\frac{n}{d}\right)^{s-1} = \frac{1}{n^{s-1}} \sigma_{s-1}(n) \end{aligned}$$

EXERCISES

- 23-1. Prove that if $f(n) = O(1)$ and $s > 1$, then $\sum f(n) n^{-s}$ is absolutely convergent.
- 23-2. Fill in the details of the proof of Theorem 23.3.
- 23-3. Find an asymptotic expression for $\sum_{n \leq x} \mu(n) n^{-4}$.
- 23-4. If s and k are suitably restricted, prove
- $\zeta^2(s) = \sum \tau(n) n^{-s}$
 - $\zeta(s) \zeta(s-1) = \sum \sigma(n) n^{-s}$
 - $\zeta(s) \zeta(s-k) = \sum \sigma_k(n) n^{-s}$
 - $\zeta(s-1)/\zeta(s) = \sum \varphi(n) n^{-s}$
- 23-5. Consider again the asymptotic formula of Exercise 20-3. Show that the limit of the left side as $x \rightarrow \infty$ is the infinite series $\zeta(2)$, and the limit of the right side is the constant K of Exercise 20-3.

24. Squarefree Integers

We have shown that $\mu^2(n)$ is 1 if n is squarefree, 0 otherwise. Hence $\sum_{n \leq x} \mu^2(n)$ is the number of squarefree integers not exceeding x , and we will now study this function. We first notice that

$$\mu^2(n) = \sum_{d^2|n} \mu(d)$$

because n can be written uniquely in the form $n = N^2q$, q squarefree. Hence,

$$\sum_{d^2|n} \mu(d) = \sum_{d|N} \mu(d) = \varepsilon(N) = \mu^2(n).$$

Theorem 24.1.

$$\sum_{n \leq x} \mu^2(n) = \frac{6x}{\pi^2} + O(\sqrt{x})$$

Proof. In the summations below, we are summing (in two ways) over the lattice points, whose first coordinate is a square, in the region bounded by $u = 1/2$, $v = 1/2$, and $u^2v = x$.

$$\begin{aligned} \sum_{n \leq x} \mu^2(n) &= \sum_{n \leq x} \sum_{d^2|n} \mu(d) \\ &= \sum_{d \leq \sqrt{x}} \mu(d) \sum_{n \leq x/d^2} 1 \\ &= \sum_{d \leq \sqrt{x}} \mu(d) \left\{ \frac{x}{d^2} + O(1) \right\} \\ &= x \sum_{d \leq \sqrt{x}} \frac{\mu(d)}{d^2} + O(\sqrt{x}) \\ &= x \{ 1/\zeta(2) + O(1/\sqrt{x}) \} + O(\sqrt{x}) \\ &= \frac{6x}{\pi^2} + O(\sqrt{x}) \end{aligned}$$

Chapter 5

SUMS OF SQUARES

25. Sums of Four Squares

In this section we will prove that if $n \in \mathcal{Z}^+$, then there are integers x_j ($j = 1, 2, 3, 4$) such that $n = x_1^2 + x_2^2 + x_3^2 + x_4^2$; that is, every positive integer can be written as a sum of four squares. The proof is greatly simplified by noting the identity

$$\begin{aligned}
 \sum_{j=1}^4 x_j^2 \sum_{j=1}^4 y_j^2 &= (x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4)^2 \\
 &\quad + (x_1 y_2 - x_2 y_1 + x_3 y_4 - x_4 y_3)^2 \\
 &\quad + (x_1 y_3 - x_3 y_1 + x_4 y_2 - x_2 y_4)^2 \\
 &\quad + (x_1 y_4 - x_4 y_1 + x_2 y_3 - x_3 y_2)^2.
 \end{aligned}
 \tag{1}$$

Because of (1) it suffices to prove that every prime can be written as a sum of four squares. Since $2 = 1^2 + 1^2 + 0^2 + 0^2$, we may restrict our attention to odd primes.

Lemma 25.1. *If p is an odd prime, there exist $a, b \in \mathcal{Z}$ such that $0 \leq a \leq (p-1)/2$, $0 \leq b \leq (p-1)/2$, and $a^2 + b^2 + 1 \equiv 0 \pmod{p}$.*

Proof. Consider the sets

$$\begin{aligned}
 \mathcal{A} &= \left\{ a^2 : a = 0, 1, \dots, \frac{p-1}{2} \right\}, \\
 \mathcal{B} &= \left\{ -b^2 - 1 : b = 0, 1, \dots, \frac{p-1}{2} \right\}.
 \end{aligned}$$

Clearly no two a 's in $\mathcal{A}$ are congruent $\pmod{p}$, for if $a_1^2 \equiv a_2^2$, then either $a_1 - a_2 \equiv 0$ or $a_1 + a_2 \equiv 0$. If $a_1 \equiv a_2 \pmod{p}$, then $a_1 = a_2$ since $0 \leq a_j \leq (p-1)/2$; if $a_1 + a_2 \equiv 0$, then $a_1 = a_2 = 0$. Similarly, no two elements

in $\mathcal{B}$ are congruent. But in $\mathcal{A}$ and $\mathcal{B}$ there are $p + 1$ elements, so that some $a^2 \in \mathcal{A}$ is congruent (mod p) to some $-b^2 - 1 \in \mathcal{B}$. This proves the lemma.

From the lemma we have $a^2 + b^2 + 1^2 + 0^2 = kp$ for some $k > 0$; also, since

$$0 < kp = a^2 + b^2 + 1 \leq 2\left(\frac{p-1}{2}\right)^2 + 1 < p^2,$$

$k < p$. We have the following result:

There exist $x_j \in \mathcal{Z}$ ($j = 1, 2, 3, 4$) such that

$$(2) \quad kp = \sum_{j=1}^4 x_j^2,$$

$$0 < k < p, \text{ and } x_j \not\equiv 0 \pmod{p} \text{ for at least one } x_j.$$

If we can show that the least k for which (2) holds is $k = 1$, then we will have shown that every odd prime p is expressible as a sum of four squares. Let m be the least positive k for which (2) holds.

Assume $m > 1$. If m is even, then we may assume that

$$x_1 \equiv x_2 \pmod{2} \quad \text{and} \quad x_3 \equiv x_4 \pmod{2}.$$

That is, if m is even, then either none, two, or four of the x_j are even and we may choose notation so that the above congruences hold. Then

$$\left\{\frac{x_1 + x_2}{2}\right\}^2 + \left\{\frac{x_1 - x_2}{2}\right\}^2 + \left\{\frac{x_3 + x_4}{2}\right\}^2 + \left\{\frac{x_3 - x_4}{2}\right\}^2 = \frac{m}{2}p$$

is a representation of $(m/2)p$ as a sum of the squares of four *integers*; by the way m was chosen, from (2) we conclude that

$$\frac{x_1 + x_2}{2} \equiv \frac{x_1 - x_2}{2} \equiv \frac{x_3 + x_4}{2} \equiv \frac{x_3 - x_4}{2} \equiv 0 \pmod{p}.$$

But then

$$\frac{x_1 + x_2}{2} \pm \frac{x_1 - x_2}{2} \equiv 0 \pmod{p}$$

implies that $x_1 \equiv x_2 \equiv 0$; similarly, $x_3 \equiv x_4 \equiv 0 \pmod{p}$ and this contradicts (2). Hence m is odd.

Since $m \geq 3$ by assumption, we can find y_j in an absolutely least CRS (mod m), i.e., $|y_j| < m/2$, such that $y_j \equiv x_j \pmod{m}$ for $j = 1, 2, 3, 4$. Equivalently, there exist m_j such that

$$(3) \quad y_j = x_j + m_j m, \quad j = 1, 2, 3, 4.$$

Now

$$\begin{aligned}\sum_{i=1}^4 y_j^2 &= \sum x_j^2 + 2m \sum x_j m_j + m^2 \sum m_j^2 \\ &\equiv \sum x_j^2 \equiv 0 \pmod{m}.\end{aligned}$$

Therefore, for some M we have $Mm = \Sigma y_j^2$. Since $|y_j| < m/2$,

$$0 \leq Mm < 4(m/2)^2 = m^2.$$

If $M = 0$, then $y_j = 0$ for all j . But then from (3), $x_j = -m_j m$, and from (2) we get $mp = m^2 \Sigma (-m_j)^2$, or $m|p$, which is impossible. Therefore, $0 < Mm < m^2$, so $0 < M < m$.

We now have $\Sigma x_j^2 = mp$, $\Sigma y_j^2 = Mm$, and $0 < M < m < p$. Let

$$w_1 = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

$$w_2 = x_1 y_2 - x_2 y_1 + x_3 y_4 - x_4 y_3$$

$$w_3 = x_1 y_3 - x_3 y_1 + x_4 y_2 - x_2 y_4$$

$$w_4 = x_1 y_4 - x_4 y_1 + x_2 y_3 - x_3 y_2$$

and using (1) we find that

$$(4) \quad \sum_{j=1}^4 w_j^2 = Mm^2 p.$$

Notice that $w_1 = \Sigma x_j y_j = \Sigma x_j (x_j + m_j m) = \Sigma x_j^2 + m \Sigma x_j y_j \equiv 0 \pmod{m}$. In a similar manner, $w_2 \equiv w_3 \equiv w_4 \equiv 0 \pmod{m}$. We write $w_j = W_j m$ and from (4) get $\Sigma W_j^2 = Mp$. From (2) we conclude $W_j \equiv 0 \pmod{p}$, $j = 1, 2, 3, 4$. Say $W_j = D_j p$; then $Mp = \Sigma W_j^2 = p^2 \Sigma D_j^2$, from which we have $p|M$, and this is impossible. Hence $m = 1$ and every prime is the sum of four squares.

Together with (1), this proves the following theorem.

Theorem 25.1. *Every positive integer can be written as the sum of four squares.*

Example. To express $231 = 3 \cdot 7 \cdot 11$ as a sum of four squares, we use (1) as follows:

$$\begin{aligned}3 \cdot 7 &= (1^2 + 1^2 + 1^2 + 0^2)(2^2 + 1^2 + 1^2 + 1^2) \\ &= (2 + 1 + 1 + 0)^2 + (1 - 2 + 1 - 0)^2 \\ &\quad + (1 - 2 + 0 - 1)^2 + (1 - 0 + 1 - 1)^2 \\ &= 4^2 + 0^2 + (-2)^2 + 1^2.\end{aligned}$$

Then again by (1) we have

$$\begin{aligned}(3 \cdot 7) \cdot 11 &= (4^2 + 2^2 + 1^2 + 0^2)(3^2 + 1^2 + 1^2 + 0^2) \\ &= (12 + 2 + 1 + 0)^2 + (4 - 6 + 0 - 0)^2 \\ &\quad + (4 - 3 + 0 - 0)^2 + (0 - 0 + 2 - 1)^2 \\ &= 15^2 + (-2)^2 + 1^2 + 1^2.\end{aligned}$$

EXERCISES

- 25-1. Express 105, 385, and 625 as sums of four squares.
 25-2. In the example given in this section, we showed that 3 and 11 can be written as a sum of three squares. Prove that 7 *cannot* be expressed as a sum of three squares.
 25-3. Prove if $n \equiv -1 \pmod{8}$, then n cannot be written as a sum of three squares. [Hint: If x, y, z are in a CRS $\pmod{8}$, consider the possible values for $x^2 + y^2 + z^2 \pmod{8}$.]

26. Sums of Two Squares

If there are integers x and y such that $n = x^2 + y^2$ and $(x, y) = 1$, then we say n has a *primitive representation* as a sum of two squares; if $(x, y) > 1$, we call $n = x^2 + y^2$ an *imprimitive representation*.

It is obvious that not every integer can be written as a sum of two squares. For example, $3 = x^2 + y^2$ has no solution x, y in integers. We will first determine necessary and sufficient conditions for $n \in \mathcal{Z}^+$ to be represented as a sum of two squares. When n can be so represented, we will determine the number of ways in which this can be done.

Lemma 26.1. Suppose $n \in \mathcal{Z}^+$. If there is a prime q such that $q|n$ and $q \equiv 3 \pmod{4}$, then n has no primitive representation as a sum of two squares.

Proof. Suppose there exists a prime q such that $q|n$, $q \equiv 3 \pmod{4}$, and $n = x^2 + y^2$ is a primitive representation. Evidently, $q \nmid x$ and $q \nmid y$, because if, say, $q|x$, then $q|(n - x^2) = y^2$, so $q|y$, which contradicts our assumption that $(x, y) = 1$.

Since $(x, q) = 1$, we know by Corollary 12.2 that there is some u such that $y \equiv ux \pmod{q}$. Then

$$x^2(1 + u^2) = x^2 + u^2x^2 \equiv x^2 + y^2 = n \equiv 0 \pmod{q},$$

so that $1 + u^2 \equiv 0$, or $u^2 \equiv -1 \pmod{q}$. Thus $(-1/q) = 1$. But q is of the form $q \equiv 3 \pmod{4}$ and by Euler's criterion

$$(-1/q) = (-1)^{(q-1)/2} = -1.$$

Therefore, n has no primitive representation.

Theorem 26.1. *If q is prime, $q \equiv 3 \pmod{4}$, $q^\alpha \nmid n$, α odd, then n has neither primitive nor imprimitive representations as a sum of two squares.*

Proof. Suppose $n = x^2 + y^2$ and $(x, y) = d$. Say $q^\beta \parallel d$, $\beta \geq 0$. Let $x = Xd$, $y = Yd$, with $(X, Y) = 1$. We have $n = d^2(X^2 + Y^2) = d^2N$, say. Since $q^{\alpha-2\beta} \nmid N$, we have a contradiction because $\alpha - 2\beta$ is odd, hence ≥ 1 , and $q \mid N$, $N = X^2 + Y^2$ is a primitive representation, which is impossible by the preceding lemma.

The problem of writing any n as a sum of two squares may be facilitated by using the identity

$$(5) \quad (x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1y_1 + x_2y_2)^2 + (x_1y_2 - x_2y_1)^2,$$

which tells us that if n_1 and n_2 are each representable as a sum of two squares, then so is n_1n_2 . If n is any positive integer, suppose the canonical form is

$$n = 2^\alpha \prod p_i^{\alpha_i} \prod q_j^{\beta_j}$$

where $p_i \equiv 1 \pmod{4}$ and $q_j \equiv 3 \pmod{4}$. We know from Theorem 26.1 that β_j must be even for all j if n is to have a representation; indeed, if β_j is even, then $q_j^{\beta_j} = \{q_j^{(1/2)\beta_j}\}^2 + 0^2$ is a representation of $q_j^{\beta_j}$, and by using (5) we can write $\prod q_j^{\beta_j}$ as a sum of two squares when the β_j are even. Also $2 = 1^2 + 1^2$ together with (5) shows that every power of 2 is a sum of two squares. Now we will prove that if $p \equiv 1 \pmod{4}$, then $p = x^2 + y^2$ has a solution $x, y \in \mathcal{Z}$. For this, we require the following lemma.

Lemma 26.2. *If z is real and n is a positive integer, then there are integers a and b such that*

$$\left| z - \frac{a}{b} \right| \leq \frac{1}{b(n+1)}, \quad 1 \leq b \leq n.$$

Proof. Consider the numbers $\alpha_k = kz - [kz]$, $k = 0, 1, \dots, n$. Obviously, $0 \leq \alpha_k < 1$ for $k = 0, 1, \dots, n$. By renaming, if necessary, call these $n+1$ numbers β_k such that $0 \leq \beta_0 \leq \beta_1 \leq \dots \leq \beta_n$, and consider the differences

$$\beta_1 - \beta_0, \beta_2 - \beta_1, \dots, \beta_n - \beta_{n-1}, \beta_0 - \beta_n + 1.$$

Each of these $n+1$ numbers is non-negative, and

$$(\beta_1 - \beta_0) + (\beta_2 - \beta_1) + \dots + (\beta_0 - \beta_n + 1) = 1;$$

therefore, at least one of them is $\leq 1/(n+1)$. But each of these, and in particular the one which is $\leq 1/(n+1)$, has the form $\beta_k - \beta_j + r$, where $k \neq j$, $0 \leq j \leq n$, $0 \leq k \leq n$, and r is 0 or 1. Hence

$$\frac{1}{n+1} \geq \beta_k - \beta_j + r = (k-j)z + [jz] - [kz] + r \geq 0.$$

If $k > j$, take $b = k - j$ and $a = [kz] - [jz] - r$, and if $k < j$, take $b = j - k$ and $a = [jz] - [kz] + r$. In either case, $|bz - a| \leq 1/(n + 1)$. Clearly, since j, k are distinct non-negative integers not exceeding n , we have $1 \leq b = |k - j| \leq n$, and $|z - a/b| \leq 1/b(n + 1)$.

Now if p is a prime, $p \equiv 1 \pmod{4}$, we use Lemma 26.2 to show that p is the sum of two squares. Since -1 is a quadratic residue of p , there is some u such that $u^2 \equiv -1 \pmod{p}$. In the above lemma, let $z = -u/p$ and $n = \lfloor \sqrt{p} \rfloor$. There are $a, b \in \mathcal{Z}$ such that

$$\left| -\frac{u}{p} - \frac{a}{b} \right| \leq \frac{1}{b(\lfloor \sqrt{p} \rfloor + 1)} < \frac{1}{b\sqrt{p}}, \quad 0 < b \leq \lfloor \sqrt{p} \rfloor \leq \sqrt{p}.$$

Let $c = ub + ap$, so that

$$|c| = |(ub + ap)| = \left| -\frac{u}{p} - \frac{a}{b} \right| |pb| < \sqrt{p}.$$

Therefore, $0 < b^2 + c^2 < (\sqrt{p})^2 + (\sqrt{p})^2 = 2p$. But $c \equiv ub \pmod{p}$, and $b^2 + c^2 \equiv b^2(1 + u^2) \equiv 0 \pmod{p}$. Since $0 < b^2 + c^2 < 2p$, we must have $p = b^2 + c^2$.

Together with our previous results this proves Theorem 26.2.

Theorem 26.2. Suppose

$$n = 2^\alpha \prod_{i=1}^r p_i^{\alpha_i} \prod_{j=1}^s q_j^{\beta_j}, \quad \alpha \geq 0, \quad \alpha_i > 0, \quad \beta_j > 0,$$

$p_i \equiv 1 \pmod{4}$, $q_j \equiv 3 \pmod{4}$. The equation $n = x^2 + y^2$ has an integral solution x, y if and only if β_j is even for $j = 1, \dots, s$.

Example. For $n = 2^3 \cdot 5 \cdot 13 \cdot 11^2$, we note that $2^3 = 2^2 + 2^2$, $5 = 2^2 + 1^2$, $13 = 2^2 + 3^2$, and $11^2 = 11^2 + 0^2$. Now by successive applications of (5) we have

$$\begin{aligned} 2^3 \cdot 5 &= (2^2 + 2^2)(2^2 + 1^2) = (4 + 2)^2 + (2 - 4)^2 = 6^2 + 2^2, \\ (2^3 \cdot 5) \cdot 13 &= (6^2 + 2^2)(2^2 + 3^2) = (12 + 6)^2 + (18 - 4)^2 = 18^2 + 14^2, \\ (2^3 \cdot 5 \cdot 13) \cdot 11^2 &= (18^2 + 14^2)(11^2 + 0^2) = (198 + 0)^2 + (0 - 154)^2. \end{aligned}$$

EXERCISES

- 26-1.* Show that $2^{2\alpha} = (\pm 2^\alpha)^2 + 0^2$ and $2^{2\alpha+1} = (\pm 2^\alpha)^2 + (\pm 2^\alpha)^2$ are the only representations as sums of two squares for powers of 2.
- 26-2. Express 45, 325, and 5929 as a sum of two squares.
- 26-3. Prove that for every $\beta \in \mathcal{Z}^+$, the equation $3^{2\beta} = x^2 + y^2$ has only the four solutions $x = \pm 3^\beta$, $y = 0$ and $x = 0$, $y = \pm 3^\beta$.

27. Number of Representations

Exercise 26-1 shows that a power of 2 may be written as a sum of two squares in four ways, namely, $2^{2\alpha} = (2^\alpha)^2 + 0^2 = (-2^\alpha)^2 + 0^2 = 0^2 + (2^\alpha)^2 = 0^2 + (-2^\alpha)^2$. This is a special case of the general problem which we now consider: If $n \in \mathcal{Z}^+$, in how many ways can n be written as a sum of two squares? Suppose

$$n = 2^\alpha \prod_{k=1}^r p_k^{\beta_k} \prod_{j=1}^s q_j^{\gamma_j}$$

where $\alpha \geq 0$, $p_k \equiv 1 \pmod{4}$, $q_j \equiv 3 \pmod{4}$. For convenience we write $n_1 = \prod p_k^{\beta_k}$ and $n_2 = \prod q_j^{\gamma_j}$, so we have $n = 2^\alpha n_1 n_2$.

Theorem 27.1. *If n is as above, then the number $N(n)$ of ways in which n can be written as a sum of two squares is*

$$N(n) = \begin{cases} 4\tau(n_1) & \text{if } n_2 \text{ is a square;} \\ 0 & \text{if } n_2 \text{ is not a square.} \end{cases}$$

The proof of this theorem will be postponed until we investigate some of the properties of the Gaussian integers in the next section.

28. The Gaussian Integers

A *Gaussian integer* is a complex number of the form $a + bi$, where $a, b \in \mathcal{Z}$ and $i^2 = -1$. Let $\mathcal{G}$ denote the set of all Gaussian integers. Two Gaussian integers $a + bi$ and $c + di$ are *equal* if and only if $a = c$ and $b = d$. Addition, subtraction, and multiplication in $\mathcal{G}$ are defined just as these operations are defined on the set of all complex numbers, namely

$$\begin{aligned} (a + bi) \pm (c + di) &= (a \pm c) + (b \pm d)i \\ (a + bi)(c + di) &= (ac - bd) + (ad + bc)i. \end{aligned}$$

Evidently, if $\alpha, \beta \in \mathcal{G}$, then $\alpha \pm \beta$ and $\alpha\beta \in \mathcal{G}$. Also, the set $\mathcal{Z}$ of rational integers is a subset of $\mathcal{G}$, and the multiplicative identity in $\mathcal{G}$ is the (Gaussian) integer 1. A Gaussian integer α is called a *unit* in $\mathcal{G}$ if and only if there exists $\beta \in \mathcal{G}$ such that $\alpha\beta = 1$.

A mapping D , called the *norm*, from $\mathcal{G}$ into the set of non-negative rational integers is defined by $D(a + bi) = a^2 + b^2$. This function has the property that $D(\alpha\beta) = D(\alpha)D(\beta)$ for all $\alpha, \beta \in \mathcal{G}$ (the proof is left as an exercise). Thus it is easy to see that α is a unit in $\mathcal{G}$ only if $D(\alpha)|D(1) = 1$, hence only if $D(\alpha) = 1$, since $D(\alpha) \geq 0$. But if $\alpha = a + bi$, $D(\alpha) = a^2 + b^2 = 1$ implies either $a = \pm 1$

and $b = 0$, or $a = 0$ and $b = \pm 1$. Therefore, if α is a unit, α must be of the form ± 1 or $\pm i$. Clearly, these four numbers are units and hence the only units in $\mathcal{G}$. We notice that α is a unit if and only if $D(\alpha) = 1$, and the units are all of the form i^n , $n \in \mathcal{L}$.

Two integers $\alpha, \beta \in \mathcal{G}$ are called *associates* if $\alpha = \beta\gamma$ for some unit γ . Thus α and β are associates if and only if $\alpha = \pm\beta$ or $\alpha = \pm\beta i$. If $\alpha, \beta \in \mathcal{G}$ we say α *divides* β , and write $\alpha|\beta$, if and only if there exists $\gamma \in \mathcal{G}$ such that $\alpha\gamma = \beta$. Clearly a unit is a divisor of every Gaussian integer. A non-zero, non-unit element $\alpha \in \mathcal{G}$ is called a *prime* if the only divisors of α are units or associates of α .

If $\alpha \in \mathcal{G}$ and $D(\alpha)$ is a rational prime, then α is prime in $\mathcal{G}$, for if $\alpha = \beta\gamma$ is any factorization, then $D(\alpha) = D(\beta)D(\gamma)$ and since $D(\alpha)$ is a prime in $\mathcal{L}$, either $D(\beta)$ or $D(\gamma) = 1$. We will show later that there are primes $\delta \in \mathcal{G}$ such that $D(\delta) \in \mathcal{L}$ is composite.

Example. The Gaussian integers $1 + i$, $2 + i$, $1 - i$, $2 - i$ are primes in $\mathcal{G}$ since their norms are prime in $\mathcal{L}$. The primes $1 + i$ and $1 - i$ are associated, since $(1 + i)(-i) = 1 - i$.

Every element of $\mathcal{G}$ which is not zero and not a unit can be factored into primes, as we now show. Let $\alpha \neq 0$, α not a unit. If α is a prime, we are through. If α is not a prime, then α has a representation of the form $\alpha = \beta\gamma$ with $D(\beta) > 1$ and $D(\gamma) > 1$. Evidently we also have $D(\beta) < D(\alpha)$ and $D(\gamma) < D(\alpha)$ since $D(\alpha) = D(\beta)D(\gamma)$. Now if β and γ are not primes, we may continue the factorization process. Since the norms of the factors form a strictly decreasing sequence of positive integers, the process must terminate when every further factorization involves a factor whose norm is 1.

Thus, a non-zero, non-unit in $\mathcal{G}$ may be written as a product of primes in $\mathcal{G}$. To show that this factorization is essentially (except for order and the occurrence of units and associates) unique, we require the Gaussian analogue of the division algorithm for rational integers.

Theorem 28.1. *If $\alpha, \beta \in \mathcal{G}$, $\beta \neq 0$, then there exist $\lambda, \theta \in \mathcal{G}$ such that $\alpha = \beta\lambda + \theta$, $D(\theta) < D(\beta)$.*

Proof. The proof is greatly facilitated by operating in the set of all complex numbers, rather than in the set $\mathcal{G}$. Recall that if γ is a complex number, γ has a unique representation of the form $\gamma = x + yi$ with x, y real. Also, the *conjugate* $\bar{\gamma}$ of γ is the complex number $\bar{\gamma} = x - yi$. Equality, addition (and subtraction), and multiplication of complex numbers are defined just as these concepts are defined in $\mathcal{G}$, and if $\gamma = x + yi$, $\delta = u + vi$, $\delta \neq 0$ are complex numbers, then division γ/δ is defined by

$$\frac{\gamma}{\delta} = \frac{xu + yv}{u^2 + v^2} + \frac{yu - xv}{u^2 + v^2}i.$$

Finally, the *modulus* $|\gamma|$ of a complex number γ is defined by $|\gamma| = (\gamma\bar{\gamma})^{1/2}$. In particular, notice that if the complex number α is also a Gaussian integer, then $D(\alpha) = |\alpha|^2 = \alpha\bar{\alpha}$.

Now suppose $\alpha, \beta \in \mathcal{G}$, $\beta \neq 0$. Then α/β is defined and is some complex number, say $\alpha/\beta = x + yi$ where x and y are rational. We now choose rational integers m and n such that $|x - m| \leq 1/2$, $|y - n| \leq 1/2$. This can be done by using Lemma 26.2 with the n of that lemma equal to 1.

Take $\lambda = m + ni$ and $\theta = \alpha - \beta\lambda$. We must show that $D(\theta) < D(\beta)$. But we have

$$\begin{aligned} |\theta| &= |\alpha - \beta\lambda| = |\beta| \left| \frac{\alpha}{\beta} - \lambda \right| = |\beta| |(x - m) + (y - n)i| \\ &= |\beta| \{(x - m)^2 + (y - n)^2\}^{1/2} \\ &\leq |\beta| \{(\tfrac{1}{2})^2 + (\tfrac{1}{2})^2\}^{1/2} < |\beta|. \end{aligned}$$

Hence $D(\theta) = |\theta|^2 < |\beta|^2 = D(\beta)$.

With this theorem we can define a Euclidean algorithm in $\mathcal{G}$ as follows. Suppose $\alpha, \beta \in \mathcal{G}$, $\beta \neq 0$. We find integers in $\mathcal{G}$ such that

$$\begin{aligned} \alpha &= \beta\lambda_1 + x_1, & D(x_1) < D(\beta) \\ \beta &= x_1\lambda_2 + x_2, & D(x_2) < D(x_1) \\ &\dots \\ x_{n-2} &= x_{n-1}\lambda_n + x_n, & D(x_n) < D(x_{n-1}) \\ x_{n-1} &= x_n\lambda_{n+1} + x_{n+1}, & D(x_{n+1}) = 0. \end{aligned}$$

The process must terminate with some remainder x_{n+1} such that $D(x_{n+1}) = 0$, since the norms form a strictly decreasing sequence of non-negative rational integers. The remainder x_n is a common divisor of α and β , and every common divisor divides x_n . Hence, we call x_n a *greatest common divisor* of α and β and write $x_n = (\alpha, \beta)$. This gcd is not unique, since any associate of x_n is also a gcd. However, any two gcd's must be associates, for suppose $\gamma = (\alpha, \beta)$ and $\delta = (\alpha, \beta)$. Then $\gamma|\delta$, say $\gamma\sigma = \delta$, and $\delta|\gamma$, say $\delta\pi = \gamma$. Therefore $\gamma = \delta\pi = \gamma\sigma\pi$, so $\sigma\pi = 1$, σ is a unit and γ and δ are associates.

The Euclidean algorithm allows us to prove that if $(\alpha, \beta) = 1$ and $\alpha|\beta\gamma$, then $\alpha|\gamma$. In addition, then, if α is a prime in $\mathcal{G}$ and $\alpha|\beta\gamma$, either $\alpha|\beta$ or $\alpha|\gamma$. This in turn yields a proof of the essential uniqueness of factorization in $\mathcal{G}$. The details are very similar to the proof of the same theorem for rational integers, and are omitted here.

Theorem 28.2. *If $\alpha \in \mathcal{G}$ is a non-zero, non-unit, then α can be written as a product of primes in $\mathcal{G}$, and the factorization is essentially unique.*

We consider now the problem of determining the primes in $\mathcal{G}$. We have already shown that $1 + i$, $1 - i$, $-1 + i$, and $-1 - i$ are all prime in $\mathcal{G}$ since the norms are 2, and evidently these primes are associates.

Suppose q is a rational prime and $q \equiv 3 \pmod{4}$. Let $q = \alpha\beta$ be any factorization of q in $\mathcal{G}$. We then have $D(q) = q^2 = D(\alpha)D(\beta)$. By unique factorization of rational integers, either $D(\alpha) = 1$, q , or q^2 . If $D(\alpha) = 1$, then α is a unit; also, if $D(\alpha) = q^2$, then β is a unit. If $D(\alpha) = D(\beta) = q$, and if $\alpha = a + bi$, then $D(\alpha) = q = a^2 + b^2$. But q has no representation as a sum of two squares. Therefore, every factorization of $q \equiv 3 \pmod{4}$ involves a unit factor, so q is prime in $\mathcal{G}$. Note that $D(q)$ is not a rational prime.

Suppose p is a rational prime, $p \equiv 1 \pmod{4}$. We recall that p can be written as a sum of squares, say $p = a^2 + b^2$. Thus, in $\mathcal{G}$ we have the factorization $p = (a + bi)(a - bi)$. Notice that $a + bi$ and $a - bi$ are not associates, for if they were, we would have $(a + bi)i^n = a - bi$ for some n , $0 \leq n \leq 3$. But $ai^n + bi^{n+1} = a - bi$ is impossible since a and b are relatively prime and non-zero. Also, the numbers $a \pm bi$ are primes in $\mathcal{G}$, since $D(a \pm bi) = p$.

The primes determined above are in fact the only primes in $\mathcal{G}$, as we will show.

Theorem 28.3. *The primes in $\mathcal{G}$ are precisely the numbers $1 + i$, the rational primes $q \equiv 3 \pmod{4}$, and the numbers $a \pm bi$, where $a^2 + b^2$ is a rational prime $p \equiv 1 \pmod{4}$, and associates of these.*

Proof. We have already shown that each of the numbers described in the theorem is prime in $\mathcal{G}$. To complete the proof that there are no others, let α be a prime of $\mathcal{G}$. Since $D(\alpha) = \alpha\bar{\alpha}$, $\alpha|D(\alpha)$. Thus, there are rational positive integers which are divisible by α . Let m be the least positive integer such that $\alpha|m$. If m is not a rational prime, then $m = n_1n_2$, with $1 < n_j < m$ ($j = 1, 2$), and $\alpha|n_1$ or $\alpha|n_2$ contrary to the way m was chosen. Thus α divides at least one rational prime. Suppose p and q are distinct rational primes such that $\alpha|p$ and $\alpha|q$. Since $(p, q) = 1$, there are $a, b \in \mathcal{Z}$ such that $pa + qb = 1$, and $\alpha|(pa + qb) = 1$, which is impossible since α is prime. Therefore, a prime $\alpha \in \mathcal{G}$ divides exactly one positive rational prime.

Let $\alpha = a + bi$ and let p be the prime in $\mathcal{Z}$ such that $\alpha\beta = p$ for some $\beta \in \mathcal{G}$. Taking norms, we have $D(\alpha)|p^2$. Hence, either $D(\alpha) = p$, or $D(\alpha) = p^2$ in which case $D(\beta) = 1$ and α is an associate of p .

Case 1, $p = 2$.

If $D(\alpha) = 2 = a^2 + b^2$, we have $a = \pm 1$, $b = \pm 1$. If $D(\alpha) = 4$, then α is an associate of $2 = (1 + i)(1 - i)$, which is not prime in $\mathcal{G}$.

Case 2, $p \equiv 1 \pmod{4}$.

If $D(\alpha) = p$, then $\alpha = \pm a \pm bi$, where $p = a^2 + b^2$; α cannot be an associate of p , because we have seen p is not prime in $\mathcal{G}$.

Case 3, $p \equiv 3 \pmod{4}$.

We know $D(\alpha) = p$ is impossible. Thus, α is an associate of p .

EXERCISES

- 28-1. Find $(1 - 3i, 3 + 5i)$; express $1 - 3i$ and $3 + 5i$ as products of primes in $\mathcal{G}$.
- 28-2. Prove that if $\alpha, \beta \in \mathcal{G}$, then $D(\alpha\beta) = D(\alpha)D(\beta)$. Show that $D(\alpha) = 0$ if and only if $\alpha = 0$. Show that $D(\bar{\alpha}) = D(\alpha)$, and that $D(n\alpha) = n^2D(\alpha)$ for all $n \in \mathcal{Z}$.
- 28-3. Prove or disprove: For all $\alpha, \beta \in \mathcal{G}$, $D(\alpha) + D(\beta) \leq D(\alpha + \beta)$.
- 28-4. Suppose p is a prime in $\mathcal{Z}$. Prove that the Legendre symbol $(-4/p)$ is
- undefined if and only if p is the product of two associated conjugate primes in $\mathcal{G}$ (then p is said to *ramify* in $\mathcal{G}$);
 - $= -1$ if and only if p stays prime in $\mathcal{G}$;
 - $= +1$ if and only if p is the product of two non-associated conjugate primes in $\mathcal{G}$ (p *splits* in $\mathcal{G}$).
- 28-5. Give an example to show that for some $\alpha, \beta \in \mathcal{G}$, $\beta \neq 0$, there exist $\lambda_1, \theta_1, \lambda_2, \theta_2 \in \mathcal{G}$ such that

$$\alpha = \beta\lambda_1 + \theta_1, \quad D(\theta_1) < D(\beta)$$

$$\alpha = \beta\lambda_2 + \theta_2, \quad D(\theta_2) < D(\beta)$$

and $D(\theta_1) \neq D(\theta_2)$.

- 28-6. Show that if $(\alpha, \beta) = 1$ and $\alpha|\beta\gamma$, then $\alpha|\gamma$. Use this to supply the details of the proof of Theorem 28.2.
- 28-7. If $\alpha, \beta \in \mathcal{G}$, $\alpha\beta \neq 0$, consider the complex numbers μ satisfying

$$\mu = \frac{\alpha\beta}{(\alpha, \beta)}, \quad \text{for some } (\alpha, \beta).$$

Prove that

- if μ satisfies the above equation, then $\mu \in \mathcal{G}$;
 - if μ_1 and μ_2 satisfy the above equation, then μ_1 and μ_2 are associates;
 - $\alpha|\mu$ and $\beta|\mu$;
 - if $\alpha|\gamma$ and $\beta|\gamma$, then $\mu|\gamma$.
- 28-8. Suppose $a, b \in \mathcal{Z}$ and not both a and b are zero. If $(a, b)_{\mathcal{G}}$ denotes their gcd in $\mathcal{G}$ and $(a, b)_{\mathcal{Z}}$ denotes their gcd in $\mathcal{Z}$, prove that $(a, b)_{\mathcal{G}} = \pm(a, b)_{\mathcal{Z}}$.

29. Proof of Theorem 27.1

We are now able to prove Theorem 27.1. Suppose $n = 2^{\alpha} n_1 n_2$ where $n_1 = \Pi p^{\beta}$, $n_2 = \Pi q^{\gamma}$, and $p \equiv 1 \pmod{4}$, $q \equiv 3 \pmod{4}$. By Theorem 26.2, the number $N(n)$ of representations of n as a sum of two squares is 0 if n_2 is not a square, so we may assume that each γ is even, say $\gamma = 2\delta$. If $n = x^2 + y^2$, then in $\mathcal{G}$, $n = (x + yi)(x - yi)$, and we may count the number of pairs x, y by considering the factorizations of $x \pm yi$ in $\mathcal{G}$. We have

$$n = i^{\mu} (1 + i)^{\alpha} (1 - i)^{\alpha} \prod_{p|n} (a + bi)^{\beta} (a - bi)^{\beta} \prod_{q|n} q^{2\delta}$$

where the factor i^{μ} ($\mu = 0, 1, 2, 3$) allows all possible associated factors and $p = a^2 + b^2 = (a + bi)(a - bi)$ for each $p \equiv 1 \pmod{4}$. If $n = (x + yi)(x - yi)$, then

$$(6) \quad (x + yi) = i^{M'} (1 + i)^{A'} (1 - i)^{A''} \Pi (a + bi)^{B'} (a - bi)^{B''} \Pi q^{D'}$$

and

$$(7) \quad (x - yi) = i^{M_1} (1 + i)^{A_1} (1 - i)^{A_2} \Pi (a + bi)^{B_1} (a - bi)^{B_2} \Pi q^{D_1}$$

where $M' + M_1 = \mu$, $A' + A_1 = \alpha = A'' + A_2$, $B' + B_1 = \beta = B'' + B_2$, and $D' + D_1 = 2\delta$.

But we can in fact say much more about these exponents. It is easy to show that if π and σ are complex numbers, then $\overline{\pi\sigma} = \bar{\pi}\bar{\sigma}$. Hence (7), as the conjugate of (6), is

$$(7') \quad x - yi = (-i)^{M'} (1 - i)^{A'} (1 + i)^{A''} \Pi (a - bi)^{B'} (a + bi)^{B''} \Pi q^{D'}.$$

By comparing (7) and (7') we see that $A'' = A_1$, $A' = A_2$, $B'' = B_1$, $B' = B_2$, $D' = D_1$. With our previous observations about these exponents, we conclude that

$$(8) \quad x + yi = i^{M'} (1 + i)^{A'} (1 - i)^{\alpha - A'} \Pi (a + bi)^{B'} (a - bi)^{\beta - B'} \Pi q^{D'}$$

$$(9) \quad x - yi = i^{M'} (1 + i)^{A'} (1 - i)^{\alpha - A'} \Pi (a + bi)^{\beta - B'} (a - bi)^{B'} \Pi q^{D'}.$$

Thus, different factorizations may be obtained by choosing $0 \leq M' \leq 3$, $0 \leq A' \leq \alpha$, $0 \leq B' \leq \beta$. However, notice that

$$\begin{aligned} (1 + i)^{A'+1} (1 - i)^{\alpha - (A'+1)} &= (1 + i)^{A'} (1 - i)^{\alpha - A'} (1 + i)(1 - i)^{-1} \\ &= (1 + i)^{A'} (1 - i)^{\alpha - A'}. \end{aligned}$$

Hence, all the variations obtained by choosing different values for A' may be obtained by choosing different M' . That this is *not* the case with the factors $a \pm bi$ is seen as follows. Suppose for some choice of M, M', B, B' we have

$$i^M (a + bi)^B (a - bi)^{\beta - B} = i^{M'} (a + bi)^{B'} (a - bi)^{\beta - B'}.$$

By uniqueness of factorization in $\mathcal{A}$ either $B = B'$, or $(a + bi)^B$ is an associate of $(a - bi)^{B-B'}$. But we have already seen that $a + bi$ and $a - bi$ are non-associates.

Therefore, distinct expressions in (8) and (9) are found by taking $0 \leq M' \leq 3$, and $0 \leq B' \leq \beta$. Then $N(n) = 4\Pi(\beta + 1) = 4\tau(n_1)$, as was to be proved.

30. Restatement of Theorem 27.1

Theorem 27.1 is sometimes stated in another form. We define the following arithmetic functions. Let

$$\tau_k(n) = \sum_{\substack{d|n \\ d \equiv k \pmod{4}}} 1 \quad \text{for } k = 1 \text{ and } 3.$$

Thus, $\tau_1(n)$ is the number of divisors of n which are of the form $4m + 1$, and $\tau_3(n)$ is the number of divisors of the form $4m + 3$.

Theorem 30.1.

$$N(n) = 4\{\tau_1(n) - \tau_3(n)\}$$

The proof of this theorem requires some knowledge of the functions τ_1 and τ_3 . Notice that neither of these functions is multiplicative; for example, $\tau_1(3) = \tau_1(7) = \tau_3(3) = \tau_3(7) = 1$, but $\tau_1(21) = 2 = \tau_3(21)$. Although not multiplicative, the functions do have these interesting properties: If $(a, b) = 1$, then

$$(10) \quad \tau_1(ab) = \tau_1(a)\tau_1(b) + \tau_3(a)\tau_3(b)$$

$$(11) \quad \tau_3(ab) = \tau_1(a)\tau_3(b) + \tau_1(b)\tau_3(a).$$

Equation (10) holds because every divisor d of ab may be written uniquely as $d = AB$, where $A|a$, $B|b$; then $d \equiv 1 \pmod{4}$ if and only if $A \equiv B \equiv 1 \pmod{4}$ or $A \equiv B \equiv 3 \pmod{4}$, and $d \equiv 3 \pmod{4}$ if and only if $A \equiv 1$, $B \equiv 3$ or $A \equiv 3$, $B \equiv 1 \pmod{4}$. But there are $\tau_1(ab)$ ways to choose $d \equiv 1$, while there are $\tau_1(a)\tau_1(b)$ ways to choose $A \equiv B \equiv 1$ plus $\tau_3(a)\tau_3(b)$ ways to choose $A \equiv B \equiv 3 \pmod{4}$, and this proves (10). In a similar manner (11) is proved. The remainder of the proof of Theorem 30.1 is easy to complete (see Exercises below).

EXERCISES

30-1. Suppose $n = 2^a n_1 n_2$, where the canonical form of n_1 contains only primes $p \equiv 1 \pmod{4}$ and n_2 contains only primes $q \equiv 3 \pmod{4}$. Use (10) and (11) to prove that $\tau_k(n) = \tau_k(n_1 n_2)$ for $k = 1, 3$.

- 30-2. Define the arithmetic function F by $F(n) = \tau_1(n) - \tau_3(n)$. Use Exercise 30-1 to prove that F is multiplicative.
- 30-3. With n as in Exercise 30-1, and F as in Exercise 30-2, prove $F(2^a) = 1$, $F(n_1) = \tau(n_1)$, and $F(n_2) = 1$ or 0 according as n_2 is a square or not. Then use Exercise 30-2 to conclude that $F(n)$ is $\tau(n_1)$ if n_2 is a square, 0 otherwise. (*Hint*: Use Theorem 8.1.)

Chapter 6

CONTINUED FRACTIONS, FAREY SEQUENCES, THE PELL EQUATION

IN THIS CHAPTER we will study continued fractions, Farey sequences, and the Pell equation. Each of these terms will be defined later, but roughly speaking, continued fractions are certain numbers written in the form

$$x_0 + \frac{1}{x_1 + \frac{1}{x_2 + \frac{1}{x_3 + \frac{1}{x_4 + \frac{1}{x_5 + \frac{1}{x_6 + \frac{1}{x_7 + \frac{1}{x_8 + \frac{1}{x_9}}}}}}}}}}$$

where the x_i are appropriately restricted real numbers; Farey sequences are certain ordered collections of fractions between 0 and 1; and the Pell equation is the Diophantine equation $x^2 - Dy^2 = 1$, $D \in \mathcal{L}$. We will observe some interrelations of the three topics, and we conclude the chapter with some results concerning the approximation of real numbers by rationals.

31. Finite Continued Fractions

Definition. Let $x_0, x_1, \dots, x_n$ be real numbers with $x_i \neq 0$ ($i = 1, \dots, n$). A finite continued fraction with j entries ($j = 0, 1, \dots, n$) will be denoted $\|x_0, \dots, x_j\|$ and is defined inductively by

$$\begin{aligned}\|x_0\| &= x_0, \\ \|x_0, x_1\| &= x_0 + x_1^{-1}, \\ \|x_0, x_1, \dots, x_j\| &= \|x_0, \|x_1, \dots, x_j\|\|, \quad 1 < j \leq n.\end{aligned}$$

If $x_0 \in \mathcal{L}$ and $x_i \in \mathcal{L}^+$ ($i = 1, \dots, n$), the continued fraction $\|x_0, x_1, \dots, x_n\|$ will be called simple. The number x_j ($j = 0, 1, \dots, n$) is called the j^{th} partial quotient of the continued fraction.

Examples.

$$(a) \quad \|1, 2, 3\| = 1 + \|2, 3\|^{-1} = 1 + (2 + 3^{-1})^{-1} = \frac{10}{7}$$

$$\begin{aligned} (b) \quad \|-2, 1, 2, 3\| &= -2 + \|1, 2, 3\|^{-1} = \frac{-13}{10} \\ &= -2 + \|1 + \|2, 3\|^{-1}\|^{-1} \\ &= \|-2, 1 + \|2, 3\|^{-1}\| \\ &= \|-2, 1, \|2, 3\|\| \end{aligned}$$

$$\begin{aligned} (c) \quad \|-2, 2, 1\| &= -2 + \|2, 1\|^{-1} \\ &= -2 + 3^{-1} = \|-2, 3\| \end{aligned}$$

Example (b) above illustrates the following theorem for $j = 1$, $n = 3$.

Theorem 31.1.

$$\|x_0, x_1, \dots, x_n\| = \|x_0, \dots, x_j, \|x_{j+1}, \dots, x_n\|\|, \quad 0 \leq j < n, \quad n \geq 1$$

Proof. For $n = 1$ the theorem is obviously true for all j , $0 \leq j < n = 1$. Suppose the theorem has been proved up to and including $n = N - 1 \geq 1$ (this supposition is referred to below as the primary inductive assumption). We will now show the theorem is true for N (thus completing the induction on n) by using induction on j . With $j = 0$, the result is immediate from the definition of a continued fraction. Suppose the result has been established for j , $0 \leq j < t < N$ (this is the secondary inductive assumption). Then

$$\begin{aligned} \|x_0, \dots, x_N\| &= \|x_0, \dots, x_{t-1}, \|x_t, \dots, x_N\|\|, \quad \text{by the secondary assumption} \\ &\quad \text{with } j = t - 1, \\ &= \left\| x_0, \dots, x_{t-1}, \|x_t, \|x_{t+1}, \dots, x_N\|\| \right\|, \quad \text{from the definition,} \\ &= \|x_0, \dots, x_{t-1}, x_t, \|x_{t+1}, \dots, x_N\|\|, \quad \text{by the primary assumption} \\ &\quad \text{with } n = t < N. \end{aligned}$$

The proof is complete.

An interesting special case of Theorem 31.1 occurs as follows.

Corollary 31.1. *If $x_n \neq 1$, $n \geq 1$, then*

$$\|x_0, \dots, x_{n-1}, x_n - 1, 1\| = \|x_0, \dots, x_{n-1}, x_n\|.$$

Proof. Since $\|x_n - 1, 1\| = x_n$,

$$\begin{aligned}\|x_0, \dots, x_{n-1}, x_n - 1, 1\| &= \|x_0, \dots, x_{n-1}, \|x_n - 1, 1\|\| \\ &= \|x_0, \dots, x_{n-1}, x_n\|.\end{aligned}$$

From the corollary it is clear that any number which has a finite continued fraction representation has at least two such representations, as illustrated in Example (c). If a continued fraction is not simple, it is evident that there is an unlimited number of non-trivial possible changes in the x_j which do not change the value of the continued fraction. However, if the fraction is simple, then there is only the one change of the type in Corollary 31.1 which does not change the value. The situation is described more completely in Theorem 31.2.

Theorem 31.2. *If $\|a_0, \dots, a_n\|$ is simple, then it has a unique rational value. If r is any rational number, then there are exactly two finite simple continued fractions with value r , and exactly one of these continued fractions has its last partial quotient equal to 1.*

Proof. The first statement of the theorem is easily proved by induction on the number of partial quotients. If every simple continued fraction with n partial quotients is rational, and if $\|a_0, \dots, a_n\|$ is any continued fraction with $n + 1$ partial quotients, then $\|a_0, \dots, a_n\| = \|a_0, \|a_1, \dots, a_n\|\|$ is rational.

We will now show that if r is rational, then r can be represented by a finite simple continued fraction, and that if $\|a_0, \dots, a_m\| = \|b_0, \dots, b_n\|$ are two equal simple continued fractions with $a_m > 1$ and $b_n > 1$, then $m = n$ and $a_i = b_i$ for $i = 0, 1, \dots, n$.

If r is rational, define the sequence of integers $a_0, a_1, \dots, a_n$ as follows. Take $a_0 = [r]$, the largest integer not exceeding r . Then

$$r = a_0 + \theta_0, \quad 0 \leq \theta_0 < 1.$$

If $\theta_0 = 0$, we are through; if $\theta_0 > 0$, then $1/\theta_0 > 1$ and we take $a_1 = [1/\theta_0] \geq 1$. We have

$$\frac{1}{\theta_0} = a_1 + \theta_1, \quad 0 \leq \theta_1 < 1.$$

Continue by induction. If $\theta_j \neq 0$, define $\theta_{j+1} = 1/\theta_j - [1/\theta_j]$ and take $a_{j+1} = [1/\theta_j]$; if $\theta_j = 0$, stop the process with a_j and $j = n$. This process (called the *continued fraction algorithm*) must stop after a finite number of steps, because $\theta_0 = r - a_0$ is rational. By induction, then, θ_j is the difference

of two rationals, hence is rational for every j . Say

$$\theta_j = \frac{c_j}{d_j}, \quad 0 \leq c_j < d_j, \quad (c_j, d_j) = 1.$$

Suppose $c_j \neq 0$. Then

$$\frac{d_j}{c_j} = \frac{1}{\theta_j} = a_{j+1} + \frac{c_{j+1}}{d_{j+1}}, \quad 0 \leq c_{j+1} < d_{j+1}$$

gives

$$\frac{c_j c_{j+1}}{d_{j+1}} = d_j - c_j a_{j+1} \in \mathcal{L}.$$

Therefore, $d_{j+1} | c_j c_{j+1}$, but $(d_{j+1}, c_{j+1}) = 1$ implies that $d_{j+1} | c_j$, so $d_{j+1} \leq c_j$ and hence $0 \leq c_{j+1} < c_j$ for all j . Thus, the c_j 's form a decreasing sequence of non-negative integers so for some j , $c_j = 0$.

We now show that the simple fraction $\|a_0, \dots, a_n\|$, with a_j defined as above, has the value r . We have

$$\begin{aligned} r &= a_0 + \theta_0 = \|a_0, \theta_0^{-1}\| = \|a_0, a_1 + \theta_1\| \\ &= \|a_0, \|a_1, \theta_1^{-1}\|\| = \|a_0, a_1, \theta_1^{-1}\| = \dots = \|a_0, \dots, a_{n-1}, \theta_n^{-1}\| \\ &= \|a_0, \dots, a_{n-1}, a_n + \theta_n\| \end{aligned}$$

but $\theta_n = 0$.

Finally, suppose

$$(1) \quad \|a_0, \dots, a_m\| = \|b_0, \dots, b_n\|, \quad a_m > 1, \quad b_n > 1.$$

Then

$$(2) \quad a_0 = [\|a_0, \dots, a_m\|] = [\|b_0, \dots, b_n\|] = b_0.$$

Since $a_0 = b_0$, it follows from (1) and the definition of a continued fraction that $\|a_1, \dots, a_m\| = \|b_1, \dots, b_n\|$. If $m \leq n$, we apply the same argument $m-1$ times to find $a_j = b_j$ ($j = 1, \dots, m-1$), and we have $\|a_m\| = \|b_m, \dots, b_n\|$. But $\|a_m\|$ is an integer, and $\|b_m, \dots, b_n\|$ can be an integer only if $m = n$, since $b_n > 1$. Hence $m = n$ and $a_m = b_m$.

Definition. Let $\|a_0, a_1, \dots, a_n\|$ be a finite continued fraction. For $0 \leq k \leq n$, the k^{th} convergent C_k to the given continued fraction is the continued fraction $\|a_0, \dots, a_k\|$.

We will now define two sequences $\{P_k\}$ and $\{Q_k\}$ associated with a continued fraction such that $C_k = P_k/Q_k$, $k = 0, 1, \dots, n$. For a given continued

fraction $\|a_0, a_1, \dots, a_n\|$, we define

$$(3) \quad \begin{aligned} P_0 &= a_0, & P_1 &= a_1 a_0 + 1, & P_k &= a_k P_{k-1} + P_{k-2} \quad (k \geq 2), \\ Q_0 &= 1, & Q_1 &= a_1, & Q_k &= a_k Q_{k-1} + Q_{k-2} \quad (k \geq 2). \end{aligned}$$

Obviously if $\|a_0, \dots, a_n\|$ is simple, then $P_k \in \mathcal{Z}$, $Q_k \in \mathcal{Z}^+$ ($k = 0, 1, \dots, n$), and $Q_k \geq k$. In Exercise 31-1, we outline a proof of the claim that $C_k = P_k/Q_k$, $k = 0, 1, \dots, n$.

Theorem 31.3. For $0 \leq k < n$, define

$$\Delta(k) = P_k Q_{k+1} - P_{k+1} Q_k.$$

Then $\Delta(k) = (-1)^{k+1}$.

Proof. For $k = 0$ and $k = 1$, the theorem is immediate from (3). If $\Delta(k-1) = (-1)^k$ for $k-1 \geq 1$, then

$$\begin{aligned} P_k Q_{k+1} - P_{k+1} Q_k &= P_k(a_{k+1} Q_k + Q_{k-1}) - (a_{k+1} P_k + P_{k-1}) Q_k \\ &= -\Delta(k-1) = (-1)^{k+1} \end{aligned}$$

Corollary 31.3a. If $\|a_0, a_1, \dots, a_n\|$ is simple, the convergents C_k are rational and in lowest terms.

Proof. The theorem leads to a solution of the Diophantine equation $P_k x + Q_k y = 1$, so $(P_k, Q_k) = 1$.

Corollary 31.3b.

$$C_{k+1} - C_k = \frac{(-1)^k}{Q_k Q_{k+1}}$$

Proof. This corollary is proved by dividing through the equation $\Delta(k) = (-1)^{k+1}$ by $-Q_k Q_{k+1}$.

We will now prove that the convergents oscillate about the value of the continued fraction with decreasing amplitude.

Theorem 31.4. Suppose $\alpha = \|a_0, a_1, \dots, a_n\|$ is simple. Then

$$C_0 < C_2 < C_4 < \dots \leq \alpha = C_n \leq \dots < C_5 < C_3 < C_1.$$

Proof. We have

$$\begin{aligned} C_{2k+2} - C_{2k} &= \frac{P_{2k+2}}{Q_{2k+2}} - \frac{P_{2k}}{Q_{2k}} \\ &= \frac{-a_{2k+2} \Delta(2k)}{Q_{2k} Q_{2k+2}} > 0 \end{aligned}$$

so the convergents with even subscripts form an increasing sequence. Also

$$C_{2k+1} - C_{2k-1} = \frac{-a_{2k+1} \Delta(2k-1)}{Q_{2k-1} Q_{2k+1}} < 0$$

so the convergents with odd subscripts form a decreasing sequence.

Finally we must show that any convergent C_{2k} is less than every convergent C_{2j+1} . For $2k < 2j+1$, Corollary 31.3b gives $C_{2j+1} - C_{2j} > 0$, so $C_{2j+1} > C_{2j} \geq C_{2k}$, since $2k \leq 2j$ and the even convergents are increasing. For $2j+1 < 2k$, Corollary 31.3b gives $C_{2k+1} - C_{2k} > 0$, so $C_{2k+1} > C_{2k}$; since $2j+1 < 2k+1$, we have $C_{2j+1} > C_{2k+1} > C_{2k}$.

EXERCISES

31-1. For $k = 0$ and $k = 1$, verify that

$$(4) \quad C_k = \frac{P_k}{Q_k}.$$

Assume that (4) has been proved for the k^{th} convergent ($k \geq 1$) of every continued fraction with at least $k+1$ entries, and prove (4) holds for $k+1$ by noting that $C_{k+1} = \|a_0, \dots, a_k, a_{k+1}\| = \|a_0, \dots, a_k + a_{k+1}^{-1}\|$ is the k^{th} convergent of some continued fraction.

31-2. Consider again the numbers c_j/d_j defined in the proof of Theorem 31.2.

If $r = a/b$, $(a, b) = 1$, prove that $b = d_0$, $d_j = c_{j-1}$ ($j = 1, \dots, n$), and that $d_n = c_{n-1} = 1$.

31-3. With c_j ($j = 0, 1, \dots, n$) as in Exercise 31-2, prove that c_{j+1} is the least non-negative residue of $c_{j-1} \pmod{c_j}$ for $j = 1, 2, \dots, n-1$.

31-4. Suppose r is rational, say $r = a/b$, $(a, b) = 1$. Use the Euclidean algorithm to find

$$a = ba_0 + c_0, \quad 0 < c_0 < b$$

$$b = c_0a_1 + c_1, \quad 0 < c_1 < c_0$$

...

$$c_{n-2} = c_{n-1}a_n + c_n, \quad c_n = 0.$$

Prove that these a_j are the a_j defined in the proof of Theorem 31.2, and these c_j/c_{j-1} are the c_j/d_j of Theorem 31.2.

32. Infinite Simple Continued Fractions

We now extend the concept of a continued fraction to an *infinite* simple continued fraction.

Definition. Let $a_1, a_2, a_3, \dots$ be any sequence of natural numbers. The infinite simple continued fraction $\|a_0, a_1, a_2, \dots\|$, $a_0 \in \mathcal{Z}$, is defined by

$$\|a_0, a_1, a_2, \dots\| = \lim_{n \rightarrow \infty} \|a_0, a_1, \dots, a_n\|$$

if this limit exists. If the limit exists and is equal to ρ , then we say $\|a_0, a_1, a_2, \dots\|$ has value ρ , or represents the real number ρ .

The question of the existence of the above limit is easily answered in the affirmative. In fact, the limit not only exists, but is always irrational.

Theorem 32.1. *The infinite simple continued fraction $\|a_0, a_1, \dots\|$ is always defined; that is, if $a_0 \in \mathcal{L}$ and $a_n \in \mathcal{Z}^+$ ($n = 1, 2, \dots$), then*

$$\lim_{n \rightarrow \infty} \|a_0, a_1, \dots, a_n\|$$

exists and is irrational. If ρ is any irrational number, then ρ has a representation as an infinite simple continued fraction, and the representation is unique.

Proof. As before, we let

$$\|a_0, a_1, \dots, a_n\| = C_n = \frac{P_n}{Q_n}$$

denote the n^{th} convergent. Since the proof of the oscillatory character of the convergents given in Theorem 31.4 did not depend upon the finiteness of the continued fraction, we also know that for infinite continued fractions the even convergents are monotone increasing, the odd convergents are monotone decreasing, and any even convergent is less than every odd convergent. Therefore, $\{C_{2k}\}$ is increasing and bounded above by C_1 , and so converges, say to U . Similarly, $\{C_{2k+1}\}$ is decreasing and bounded below by C_0 , and so converges, say to L . Clearly L and U must be the same, because if ε is any positive number, then L and U differ by less than ε . Choose $n \in \mathcal{Z}^+$ sufficiently large so that $1/4n^2 < \varepsilon$; then notice that $C_{2n} \leq L \leq C_{2n+1}$ (since C_{2n} is a lower bound for $\{C_{2k+1}\}$ and L is the greatest lower bound of $\{C_{2k+1}\}$) and $C_{2n} \leq U \leq C_{2n+1}$ (since C_{2n+1} is an upper bound for $\{C_{2n}\}$ and U is the least upper bound); hence

$$|L - U| \leq |C_{2n+1} - C_{2n}| = \frac{(-1)^{2n}}{Q_{2n}Q_{2n+1}} \leq \frac{1}{2n(2n+1)} < \frac{1}{4n^2} < \varepsilon.$$

(We have used here our previous remark that $Q_k \geq k$ for simple continued fractions.) Thus we see that the sequence $\{C_n\}$ of convergents converges to $L (= U)$.

To show that L is irrational, we first observe that 0 cannot be represented by an infinite simple continued fraction, because if $0 = \|x_0, x_1, \dots\|$, for example, then since the convergents oscillate about the value of the continued fraction we must have

$$x_0 < 0 < x_0 + x_1^{-1}.$$

But $x_0 < 0$ and $x_1 \in \mathcal{X}^+$, so $x_1^{-1} \leq 1$ and $x_0 + x_1^{-1} \leq 0$; therefore 0 has no infinite simple continued fraction representation. Consequently L above is irrational, for suppose the contrary. Then $L = \|a_0, a_1, \dots\|$ is rational, and L also has a finite continued fraction representation by Theorem 31.2, say $L = \|b_0, b_1, \dots, b_N\|$. By taking integral parts of $\|b_0, \dots, b_N\| = \|a_0, a_1, \dots\|$ and arguing as in the proof of Theorem 31.2, we get $b_j = a_j$ ($j = 0, 1, \dots, N$) and $0 = \|a_{N+1}, a_{N+2}, \dots\|$, which is impossible. Hence every infinite simple fraction has an irrational value.

Conversely, suppose ρ is irrational. That ρ has at least one representation as an infinite simple continued fraction is easily seen using the continued fraction algorithm. We find

$$\begin{aligned} \rho &= a_0 + \theta_0, & 0 \leq \theta_0 < 1 & \quad (a_0 = [\rho]) \\ \frac{1}{\theta_0} &= a_1 + \theta_1, & 0 \leq \theta_1 < 1 & \quad (a_1 = [1/\theta_0]) \\ &\dots & & \\ \frac{1}{\theta_n} &= a_{n+1} + \theta_{n+1}, & 0 \leq \theta_{n+1} < 1 & \quad (a_{n+1} = [1/\theta_n]) \\ &\dots & & \end{aligned}$$

Since ρ is irrational, $\theta_0 = \rho - a_0$ is irrational and so is $1/\theta_0$. Hence $\theta_1 = 1/\theta_0 - a_1$ is irrational. Suppose we have shown that θ_k is irrational; then $\theta_{k+1} = 1/\theta_k - a_{k+1}$ is irrational, so $\theta_{k+1} \neq 0$. Therefore, the process does not terminate and determines an infinite sequence of integers $a_0, a_1, a_2, \dots$. Now we show that $\rho = \|a_0, a_1, \dots\|$. We have $\rho = a_0 + \theta_0 = \|a_0, \theta_0^{-1}\| = \|a_0, a_1 + \theta_1\| = \dots = \|a_0, a_1, \dots, a_n, \theta_n^{-1}\|$ for all n . Let

$$C_n = \|a_0, \dots, a_n\| = \frac{P_n}{Q_n} \quad \text{and} \quad C'_{n+1} = \|a_0, \dots, a_n, \theta_n^{-1}\| = \frac{P'_{n+1}}{Q'_{n+1}}.$$

Then for any fixed n , the j^{th} convergent to C'_{n+1} is the j^{th} convergent to C_n if $0 \leq j \leq n$; hence $P'_j = P_j$ and $Q'_j = Q_j$ if $0 \leq j \leq n$. Therefore,

$$\rho = \frac{P'_{n+1}}{Q'_{n+1}} = \frac{\theta_n^{-1} P'_n + P'_{n-1}}{\theta_n^{-1} Q'_n + Q'_{n-1}} = \frac{\theta_n^{-1} P_n + P_{n-1}}{\theta_n^{-1} Q_n + Q_{n-1}},$$

and

$$\begin{aligned} |\rho - C_n| &= \left| \rho - \frac{P_n}{Q_n} \right| = \left| \frac{\Delta(n-1)}{(\theta_n^{-1} Q_n + Q_{n-1}) Q_n} \right| \\ &= \frac{1}{(\theta_n^{-1} Q_n + Q_{n-1}) Q_n} < \frac{1}{(a_{n+1} Q_n + Q_{n-1}) Q_n} = \frac{1}{Q_{n+1} Q_n} \leq \frac{1}{n^2}. \end{aligned}$$

Since $1/n^2 \rightarrow 0$ as $n \rightarrow \infty$, we see that $\lim_{n \rightarrow \infty} C_n = \rho$, or that $\|a_0, a_1, \dots\| = \rho$.

The uniqueness of this representation is proved by arguments similar to those used in proving the analogous result for finite simple continued fractions.

The following corollary is listed here for emphasis.

Corollary 32.1.

$$|\rho - C_n| = \left| \rho - \frac{P_n}{Q_n} \right| < \frac{1}{Q_n Q_{n+1}} < \frac{1}{Q_n^2}$$

for all $n \in \mathcal{L}^+$.

In Chapter 9 we will prove that ρ has a *periodic* infinite simple continued fraction representation if and only if ρ is a quadratic irrational, that is, ρ is an irrational root of an equation $ax^2 + bx + c = 0$, with $a, b, c \in \mathcal{L}$.

EXERCISES

- 32-1. Prove that the infinite simple continued fraction representation of an irrational number is unique.
- 32-2. Show that Lemma 26.2 is implied by Corollary 32.1 and that there are infinitely many integers x, y with the properties stated in the lemma.
- 32-3. Consider the continued fraction $\xi = \|1, 1, 1, \dots\|$. Since $\xi = 1 + \|1, 1, 1, \dots\|^{-1} = 1 + 1/\xi$, prove that $\xi = (1 + \sqrt{5})/2$.
- 32-4. With ξ as in Exercise 32-3, prove that $P_n = Q_{n+1}$, $n = 0, 1, \dots$ and that

$$Q_n = \frac{\left(\frac{1 + \sqrt{5}}{2}\right)^{n+1} - \left(\frac{1 - \sqrt{5}}{2}\right)^{n+1}}{\sqrt{5}}, \quad n = 0, 1, 2, \dots$$

This sequence $\{Q_n\}$ is called the *Fibonacci sequence*.

- 32-5.* With ξ as in Exercise 32-3, let $\xi = 1 + \theta_0$, and $\theta_n^{-1} = 1 + \theta_{n+1}$ ($n = 0, 1, 2, \dots$). Prove that $\theta_n^{-1} = \xi$ for all $n \geq 0$.
- 32-6.* With ξ as in Exercise 32-3, prove that

$$\lim_{n \rightarrow \infty} \frac{Q_n}{Q_{n+1}} = \frac{1}{\xi} = \frac{\sqrt{5} - 1}{2}.$$

With θ_n as in Exercise 32-5, prove that

$$\lim_{n \rightarrow \infty} \left(\theta_{n+1}^{-1} + \frac{Q_n}{Q_{n+1}} \right) = \sqrt{5}.$$

33. Farey Sequences

Definition. Suppose $m \in \mathbb{Z}^+$ and let

$$\mathcal{E}_m = \left\{ \frac{r}{s} : s \neq 0, r, s \in \mathbb{Z}, 0 \leq r \leq s \leq m, (r, s) = 1 \right\}.$$

The sequence obtained by arranging the elements of $\mathcal{E}_m$ in increasing order is called the Farey sequence of order m , and will be denoted by $\mathcal{F}_m$.

Examples.

$$\mathcal{F}_1 : \frac{0}{1}, \frac{1}{1}$$

$$\mathcal{F}_2 : \frac{0}{1}, \frac{1}{2}, \frac{1}{1}$$

$$\mathcal{F}_3 : \frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}$$

$$\mathcal{F}_4 : \frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}$$

The number of terms in $\mathcal{F}_m$ is $\nu \mathcal{E}_m = \sum_{d \leq m} \varphi(d) + 1$. If $m < n$, then $\mathcal{E}_m \subset \mathcal{E}_n$. If $a/b < c/d$ are in $\mathcal{E}_m$ and no other element of $\mathcal{E}_m$ is between these two, then we say $a/b < c/d$ are *successive* in $\mathcal{F}_m$, or that c/d is the *successor* of a/b in $\mathcal{F}_m$. The number $(a+c)/(b+d)$ is called the *mediant* of a/b and c/d ; we shall see that the mediant plays an interesting role in the theory of Farey sequences.

In our next theorem we give a method for finding the successor of a given term $a/b < 1$ in $\mathcal{F}_m$.

Theorem 33.1. Suppose $a/b \in \mathcal{E}_m$, $a < b$. If $u = c$, $v = d$ is the unique solution of $bu - av = 1$ satisfying $m - b < d \leq m$, then $a/b < c/d$ are successive in $\mathcal{F}_m$.

Proof. Since $(a, b) = 1$, there is a solution u_0, v_0 of the Diophantine equation $bu - av = 1$. Choose $t_0 \in \mathbb{Z}$ to be

$$t_0 = \left\lceil \frac{m - v_0}{b} \right\rceil.$$

Since $u = u_0 + at$, $v = v_0 + bt$ is a solution for all $t \in \mathbb{Z}$, it is obvious that there is one and only one solution u, v which satisfies $m - b < v \leq m$. This solution is $c = u_0 + at_0$, $d = v_0 + bt_0$, because

$$\frac{m - v_0}{b} - 1 < t_0 \leq \frac{m - v_0}{b}$$

gives

$$0 \leq m - b < v_0 + bt_0 = d \leq m.$$

Furthermore, $c = (1 + ad)/b > 0$ and

$$c = \frac{1 + ad}{b} \leq \frac{1 + am}{b} < \frac{1}{b} + m \leq m + 1$$

so $0 < c \leq m$. Since c, d is a solution of the given equation, $(c, d) = 1$ and

$$c = \frac{1 + ad}{b} < \frac{1}{b} + d$$

so $c \leq d$; hence $c/d \in \mathcal{E}_m$.

Now

$$\frac{c}{d} = \frac{bc}{bd} = \frac{1 + ad}{bd} = \frac{1}{bd} + \frac{a}{b} > \frac{a}{b}$$

so c/d occurs after a/b in $\mathcal{F}_m$. Suppose $a/b < c/d$ are not successive in $\mathcal{F}_m$. Say there is a term $x/y \in \mathcal{E}_m$ such that

$$\frac{a}{b} < \frac{x}{y} < \frac{c}{d}.$$

Then

$$\frac{c}{d} - \frac{x}{y} = \frac{cy - dx}{dy} \geq \frac{1}{dy}$$

because $cy - dx \in \mathcal{Z}^+$, since $0 < c/d - x/y$; and

$$\frac{x}{y} - \frac{a}{b} = \frac{bx - ay}{by} \geq \frac{1}{by}.$$

Therefore,

$$\begin{aligned} \frac{1}{bd} &= \frac{bc - ad}{bd} = \frac{c}{d} - \frac{a}{b} = \left(\frac{c}{d} - \frac{x}{y} \right) + \left(\frac{x}{y} - \frac{a}{b} \right) \geq \frac{1}{dy} + \frac{1}{by} \\ &= \frac{b + d}{bdy} > \frac{m}{bdy} \geq \frac{1}{bd}, \end{aligned}$$

which is a contradiction. Hence c/d is the successor of a/b .

Corollary 33.1a. *If $a/b < c/d$ are successive in $\mathcal{F}_m$, then $bc - ad = 1$.*

Corollary 33.1b. *Assume $a/b < u/v$ are successive in $\mathcal{F}_m$ and $u/v < c/d$ are successive in $\mathcal{F}_m$. Then u/v is the median $(a + c)/(b + d)$ of a/b and c/d .*

Proof. From Corollary 33.1a we have $bu - av = 1$ and $-du + cv = 1$.

Solving these equations for u, v we get

$$u = \frac{a + c}{bc - ad}, \quad v = \frac{b + d}{bc - ad},$$

so $u/v = (a + c)/(b + d)$.

We now make some remarks about the mediant $(a + c)/(b + d)$ of two terms $a/b < c/d$ which are successive in $\mathcal{F}_m$. From Corollary 33.1a it is clear that the mediant is a rational number which is between a/b and c/d , i.e.,

$$\frac{a}{b} < \frac{a + c}{b + d} < \frac{c}{d}.$$

Also, the mediant is in lowest terms; that is, $(a + c, b + d) = 1$ because the equation $(a + c)x + (b + d)y = 1$ has a solution $x = -d$, $y = c$. Consequently, $b + d > m$, otherwise the mediant is in $\mathcal{E}_m$ and $a/b < c/d$ are not successive in $\mathcal{F}_m$. These remarks suggest the following theorem.

Theorem 33.2. *Suppose $a/b < c/d$ are successive in $\mathcal{F}_m$. If $b + d = m + 1$, then the mediant $(a + c)/(b + d)$ is the only term between a/b and c/d in $\mathcal{F}_{m+1}$; conversely, if the mediant is between a/b and c/d in $\mathcal{F}_{m+1}$, then the mediant is the only term between them in $\mathcal{F}_{m+1}$ and $b + d = m + 1$. Hence, if $a/b < c/d$ are successive in $\mathcal{F}_m$, either they are also successive in $\mathcal{F}_{m+1}$ or they have exactly one term, their mediant, between them in $\mathcal{F}_{m+1}$, and the latter is the case if and only if $b + d = m + 1$.*

Proof. Assume $a/b < c/d$ are successive in $\mathcal{F}_m$. If $b + d = m + 1$, then the mediant is in $\mathcal{E}_{m+1}$. By Theorem 33.1, $u = a + c$, $v = b + d$ is the unique solution of

$$bu - av = 1, \quad m + 1 - b < b + d \leq m + 1$$

so $a/b < (a + c)/(b + d)$ are successive in $\mathcal{F}_{m+1}$. Again by Theorem 33.1, $u = c$, $v = d$ is the unique solution of

$$(b + d)u - (a + c)v = 1, \quad m + 1 - (b + d) < d \leq m + 1$$

so $(a + c)/(b + d) < c/d$ are successive in $\mathcal{F}_{m+1}$. Therefore, the mediant is the only term in $\mathcal{F}_{m+1}$ between a/b and c/d .

Now suppose the mediant is between a/b and c/d in $\mathcal{F}_{m+1}$. Since $b + d > m$ from the remarks above, we have $b + d \geq m + 1$. But the mediant is in $\mathcal{E}_{m+1}$, so $b + d \leq m + 1$, and the condition $b + d = m + 1$ is necessary. Moreover, by the sufficiency of the condition, the mediant is the only term between a/b and c/d in $\mathcal{F}_{m+1}$. This completes the proof.

Theorem 33.1 defines a recursive construction for finding the next term in $\mathcal{F}_m$ if any term less than 1 is given. Theorem 33.2 defines a recursive construction for finding $\mathcal{F}_{m+1}$ from $\mathcal{F}_m$.

EXERCISES

- 33-1. Construct $\mathcal{F}_5$ and $\mathcal{F}_6$.
 33-2. Find the successor c/d of $3/17$ in $\mathcal{F}_{17}$. Now find the term a/b such that $a/b < 3/17$ are successive in $\mathcal{F}_{17}$.
 33-3. Prove that $a/b \in \mathcal{E}_m$ if and only if $(b-a)/b \in \mathcal{E}_m$ and that $a/b \neq (b-a)/b$ except when $a = 1, b = 2$. Use these observations to prove that
- $$\sum_{x \in \mathcal{E}_m} x = \frac{1}{2} + \frac{1}{2} \sum_{d \leq m} \varphi(d) = \frac{1}{2} v \mathcal{E}_m.$$
- 33-4.* Suppose x', x'' are real numbers satisfying $0 \leq x' < x'' \leq 1$. Prove there is some $m \in \mathbb{Z}^+$ such that $\mathcal{E}_m$ contains an element e satisfying $x' < e < x''$.

34. The Pell Equation

We consider now the Diophantine equation $x^2 - Dy^2 = 1, D \in \mathbb{Z}$, which is called the Pell equation. If $D < 0$, then $x^2 - Dy^2 \geq 1$ for all $x, y \in \mathbb{Z}$ (except $x = y = 0$, which is not a solution) and the equality can hold only for $x = \pm 1, y = 0$, or for $x = 0, y = \pm 1$ if $D = -1$. If D is the square of an integer, say $D = d^2$, let $u = dy$ and the solutions must satisfy

$$(x - u)(x + u) = x^2 - Dy^2 = 1.$$

Hence $x + u = x - u = \pm 1$, and $u = 0, x = \pm 1$ are the only solutions. Therefore we see that the only interesting cases of the Pell equation arise when D is positive and not a square, hence if and only if $\sqrt{D}$ is irrational. Consequently, throughout this section we shall assume $\sqrt{D}$ is irrational.

We will say that the solution x, y is a *positive* solution of the Pell equation if and only if both x and y are positive. Clearly, all solutions will be known when all positive solutions are known. Let $\mathcal{S}$ be the set of numbers of the form $x + \sqrt{D}y$ for positive solutions x, y of the Pell equation, i.e.,

$$\mathcal{S} = \{x + \sqrt{D}y : x, y \in \mathbb{Z}^+, x^2 - Dy^2 = 1\}.$$

If $\mathcal{S} \neq \emptyset$, then $\mathcal{S}$ has a smallest element, because it happens that any two elements of $\mathcal{S}$ differ by at least 1 (see Exercise 34-1), so if $s \in \mathcal{S}$, the number of elements in $\mathcal{S}$ less than s cannot exceed the number of positive integers less than s . Thus, the elements of $\mathcal{S}$ not larger than s are finitely many, and from these we can select a minimum with no difficulty. Consequently, if the Pell equation has a positive solution, then it has a positive solution x, y for which $x + \sqrt{D}y$ is minimal; the corresponding solution x, y is called the

fundamental solution, and the minimal sum $x + \sqrt{D}y$ will be called the *fundamental factor*.

We will show that indeed there is always a positive solution (hence a fundamental solution), hence infinitely many positive solutions x, y , and that all the corresponding factors $x + \sqrt{D}y$ are positive integral powers of the fundamental factor. If x, y is a positive solution, we will call $x + \sqrt{D}y$ the *dominant factor* corresponding to the solution x, y , and we will refer to the conjugate $x - \sqrt{D}y$ as the *inferior factor* corresponding to x, y . This terminology is suggested by the following facts: If x, y is a positive solution, then $1 < x + \sqrt{D}y$; consequently, from $(x + \sqrt{D}y)(x - \sqrt{D}y) = 1$ we see that $x - \sqrt{D}y = (x + \sqrt{D}y)^{-1}$ is positive and less than 1, hence

$$0 < x - \sqrt{D}y < 1 < x + \sqrt{D}y.$$

Theorem 34.1. *There exists a positive solution of the Pell equation $x^2 - Dy^2 = 1$; hence, there exists a fundamental solution.*

Proof. By Corollary 32.1 there are infinitely many pairs $x, y \in \mathcal{Z}^+$ such that $|\sqrt{D} - x/y| < 1/y^2$, or such that $|x - \sqrt{D}y| = |\sqrt{D}y - x| < 1/y$. Since

$$x - \sqrt{D}y = |x| - |\sqrt{D}y| \leq |x - \sqrt{D}y| \leq \frac{1}{y} \leq y$$

we have $x < (1 + \sqrt{D})y$ and $x + \sqrt{D}y < (1 + 2\sqrt{D})y$. Then for every such pair x, y ,

$$|x^2 - Dy^2| = |x - \sqrt{D}y| |x + \sqrt{D}y| < \frac{1}{y} (1 + 2\sqrt{D})y = 1 + 2\sqrt{D}.$$

Thus, all of the forms $x^2 - Dy^2$ have integer values between $-(1 + 2\sqrt{D})$ and $1 + 2\sqrt{D}$, so there is an integer k , $-(1 + 2\sqrt{D}) < k < 1 + 2\sqrt{D}$, such that for *infinitely many* x, y we have $x^2 - Dy^2 = k$. Notice $k \neq 0$ since D is not a square, and if $k = 1$ we are through.

If $k \neq 1$, we partition the set of all such pairs x, y into equivalence classes by putting two pairs u, v and x, y into the same class if and only if

$$(5) \quad u \equiv x \pmod{|k|} \quad \text{and} \quad v \equiv y \pmod{|k|}.$$

Since there are infinitely many pairs and only k^2 classes, some class contains two pairs x_1, y_1 and x_2, y_2 of positive integers such that $x_1 \neq x_2$ and $y_1 \neq y_2$. Define

$$x_0 = \frac{x_1x_2 - Dy_1y_2}{k} \quad \text{and} \quad y_0 = \frac{x_1y_2 - x_2y_1}{k}.$$

Since x_1, y_1 and x_2, y_2 satisfy (5), we see that the numerators of x_0 and y_0 are $\equiv 0 \pmod{|k|}$, so $x_0 \in \mathcal{Z}$ and $y_0 \in \mathcal{Z}$. Also,

$$x_0^2 - Dy_0^2 = \frac{1}{k^2} (x_1^2 - Dy_1^2)(x_2^2 - Dy_2^2) = 1.$$

Finally, neither x_0 nor y_0 is zero, because if $x_0 = 0$, then $1 = x_0^2 - Dy_0^2 = -Dy_0^2 \leq 0$. If $y_0 = 0$, then $x_1 = x_2y_1/y_2$ and

$$k = x_1^2 - Dy_1^2 = \left(\frac{x_2y_1}{y_2}\right)^2 - Dy_1^2 = \left(\frac{y_1}{y_2}\right)^2 (x_2^2 - Dy_2^2) = \left(\frac{y_1}{y_2}\right)^2 k$$

so $y_1^2 = y_2^2$; but since y_1 and y_2 are positive, we have $y_1 = y_2$, and this contradicts the way the pairs x_1, y_1 and x_2, y_2 were chosen.

We will now observe that the product of dominant factors corresponding to two positive solutions is again a dominant factor corresponding to a positive solution, or what is the same thing, the product of two elements of $\mathcal{S}$ is in $\mathcal{S}$. (The same is true for inferior factors; the proof is left as an exercise.) More precisely, suppose x, y and u, v are positive solutions of the Pell equation. Then the pair a, b of integers defined by

$$(x + \sqrt{D}y)(u + \sqrt{D}v) = a + \sqrt{D}b$$

is a positive solution. Clearly, $a = xu + Dyv \in \mathcal{Z}^+$ and $b = xv + yu \in \mathcal{Z}^+$. Also,

$$a^2 - Db^2 = (x^2 - Dy^2)(u^2 - Dv^2) = 1,$$

which proves the remark. An obvious corollary is that if x_1, y_1 is the fundamental solution, then for every $n \in \mathcal{Z}^+$, $(x_1 + \sqrt{D}y_1)^n$ is the dominant factor of a positive solution, so there are infinitely many solutions. The remarkable thing is that the dominant factor of every positive solution has this form, which we prove next.

Theorem 34.2. *The set $\mathcal{S}$ of dominant factors corresponding to the positive solutions of the Pell equation is precisely the set of positive integral powers of the dominant fundamental factor.*

Proof. Let x_1, y_1 be the fundamental solution. We have already proved that $\{(x_1 + \sqrt{D}y_1)^n : n \in \mathcal{Z}^+\} \subset \mathcal{S}$, so all that remains to be shown is that if u, v is a positive solution, then there exists $n \in \mathcal{Z}^+$ such that

$$u + \sqrt{D}v = (x_1 + \sqrt{D}y_1)^n.$$

Suppose $u + \sqrt{D}v$ is not a power of $x_1 + \sqrt{D}y_1$. Let $\alpha = x_1 + \sqrt{D}y_1$; then α^{-1} is the inferior fundamental factor. Because of the supposition, we can find $n \in \mathcal{Z}^+$ such that

$$\alpha^n < u + \sqrt{D}v < \alpha^{n+1}$$

so that

$$(6) \quad 1 = \alpha^n \alpha^{-n} < (u + \sqrt{D}v) \alpha^{-n} < \alpha^{n+1} \alpha^{-n} = \alpha.$$

Since the product of inferior factors is an inferior factor, $\alpha^{-n} = (\alpha^{-1})^n$ is an inferior factor corresponding to, say, the positive solution a, b , and $\alpha^{-n} = a - \sqrt{D}b$. We will show that $(u + \sqrt{D}v)\alpha^{-n}$ is a dominant factor corresponding to some positive solution; hence (6) will contradict the minimality of α .

Let

$$c + \sqrt{D}d = (u + \sqrt{D}v)\alpha^{-n} = (u + \sqrt{D}v)(a - \sqrt{D}b).$$

It is easily verified by straightforward computation that $c = au - Dbv$, $d = av - bu$ is a solution of the Pell equation. It is a positive solution, because from (6), $c + \sqrt{D}d > 1$, and since c, d is a solution, the conjugate $c - \sqrt{D}d$ is positive and less than 1. Hence the sum of these factors is positive (which implies $c > 0$) and the difference $c + \sqrt{D}d - (c - \sqrt{D}d)$ is positive (which implies $d > 0$). Thus, we have our contradiction and the proof is complete.

EXERCISES

- 34-1. Suppose $\alpha, \beta \in \mathcal{S}$, say $\alpha = x + \sqrt{D}y$ and $\beta = u + \sqrt{D}v$. If $x < u$, prove $y < v$. Hence, prove that $|\alpha - \beta| > 1$.
- 34-2. Find the fundamental solution of $x^2 - 11y^2 = 1$.
- 34-3. Prove that if $\alpha, \beta \in \mathcal{S}$, then $(\alpha^{-1}\beta^{-1})^{-1} \in \mathcal{S}$. Show that this proves the product of inferior factors corresponding to positive solutions is an inferior factor corresponding to a positive solution.
- 34-4. Suppose β is a dominant factor and γ is an inferior factor. Prove
- $\beta\gamma$ is a dominant factor if and only if $\beta > \gamma^{-1}$;
 - $\beta\gamma$ is an inferior factor if and only if $\beta < \gamma^{-1}$;
 - $\beta\gamma$ does not correspond to a positive solution if and only if $\beta = \gamma^{-1}$.
- 34-5. Prove that the set

$$\mathcal{T} = \mathcal{S} \cup \{\beta^{-1} : \beta \in \mathcal{S}\} \cup \{1\}$$

is a group under multiplication. The proof may be facilitated by first noticing that

$$\mathcal{T} = \{\beta^n : \beta \in \mathcal{S}, n \in \mathbb{Z}\}.$$

- 34-6. Suppose x, y and u, v are pairs of positive integers. We write $(x, y) \mathcal{R} (u, v)$ if and only if equation (5) is satisfied. Prove $\mathcal{R}$ is an equivalence relation on the set of pairs x, y such that $x^2 - Dy^2 = k$.
- 34-7. Suppose D is a prime. If -1 is a quadratic non-residue (mod D), prove $x^2 - Dy^2 = -1$ has no solutions.
- 34-8. If $x^2 - Dy^2 = -1$ has a solution, prove it has a positive solution x_1, y_1 for which $x_1 + \sqrt{D}y_1$ is minimal. Then prove that odd powers

of $x_1 + \sqrt{D}y_1$ correspond to positive solutions of the given equation, and even powers of $x_1 + \sqrt{D}y_1$ correspond to positive solutions of the Pell equation.

35. Rational Approximations of Reals

In this section we use some of the results of the previous sections of this chapter to find rational numbers which are in some sense "close" to a given real number.

The following theorem which we saw earlier (Lemma 26.2) is proved here using only our results on Farey sequences.

Theorem 35.1. *If ξ is irrational and $n \in \mathcal{Z}^+$, there exist integers a, b such that $(a, b) = 1$ and*

$$\left| \xi - \frac{a}{b} \right| < \frac{1}{b(n+1)}, \quad 0 < b \leq n.$$

Proof. We first assume $0 < \xi < 1$. In the Farey sequence $\mathcal{F}_n$, find the unique pair of successors $a/b < c/d$ such that ξ is between them. Since these are successive, evidently their median $(a+c)/(b+d)$ is not in $\mathcal{E}_n$, and $b+d \geq n+1$. Let us assume ξ is between a/b and the median; the arguments are the same if ξ is between the median and c/d . Then

$$\left| \xi - \frac{a}{b} \right| < \left| \frac{a+c}{b+d} - \frac{a}{b} \right| = \frac{1}{b(b+d)} \leq \frac{1}{b(n+1)},$$

and since $a/b \in \mathcal{E}_n$, $0 < b \leq n$, and $(a, b) = 1$. This proves the theorem if $0 < \xi < 1$.

If ξ is not between 0 and 1, write $\xi = [\xi] + \theta$, $0 < \theta < 1$. Let x, y be the relatively prime integers such that

$$\left| \theta - \frac{x}{y} \right| < \frac{1}{y(n+1)}, \quad 0 < y \leq n.$$

Then

$$\left| \xi - \frac{y[\xi] + x}{y} \right| = \left| \xi - [\xi] - \frac{x}{y} \right| = \left| \theta - \frac{x}{y} \right| < \frac{1}{y(n+1)},$$

and we take $a = y[\xi] + x$, $b = y$. Clearly $0 < b \leq n$, and $(a, b) = (x, y) = 1$.

Theorem 35.2. *If ξ is irrational, there are infinitely many $a, b \in \mathcal{Z}$ such that $(a, b) = 1$ and*

$$\left| \xi - \frac{a}{b} \right| < \frac{1}{b^2}.$$

This was proved earlier as a corollary (32.1) to results on the convergents to an irrational ξ . Also, notice that this theorem is obtained as a weak corollary to Theorem 35.1 in this way: For $n_1 = 1$, find a', b' such that

$$\left| \xi - \frac{a'}{b'} \right| < \frac{1}{b'(n_1 + 1)} < \frac{1}{(b')^2}.$$

Now choose $n_2 \in \mathcal{Z}^+$ such that

$$\frac{1}{n_2} < \left| \xi - \frac{a'}{b'} \right|;$$

then there exist a'', b'' such that

$$\left| \xi - \frac{a''}{b''} \right| < \frac{1}{b''(n_2 + 1)} < \frac{1}{(b'')^2},$$

and since

$$\left| \xi - \frac{a''}{b''} \right| < \frac{1}{b''(n_2 + 1)} < \frac{1}{n_2} < \left| \xi - \frac{a'}{b'} \right|,$$

the pair a'', b'' is different from a', b' . Choose n_3 such that

$$\frac{1}{n_3} < \left| \xi - \frac{a''}{b''} \right|$$

and continue by induction.

Theorem 35.3. (Hurwitz.) *If ξ is irrational, there exist infinitely many $x, y \in \mathcal{Z}$ such that*

$$\left| \xi - \frac{x}{y} \right| < \frac{1}{\sqrt{5}y^2}.$$

Proof. We again prove the theorem in case $0 < \xi < 1$; the extension is exactly the same as in the proof of Theorem 35.1.

Let $n \in \mathcal{Z}^+$ and find the successive fractions $a/b < c/d$ in $\mathcal{F}_n$ which surround ξ . If we assume that neither a/b nor c/d satisfies the inequality in the theorem, then we have by assumption

$$(7) \quad \xi - \frac{a}{b} \geq \frac{1}{\sqrt{5}b^2} \quad \text{and} \quad \frac{c}{d} - \xi \geq \frac{1}{\sqrt{5}d^2}.$$

Adding these inequalities, we obtain

$$\frac{c}{d} - \frac{a}{b} \geq \frac{1}{\sqrt{5}} \left(\frac{1}{b^2} + \frac{1}{d^2} \right);$$

by using Corollary 33.1a this can be put in the form

$$(8) \quad b^2 - \sqrt{5}bd + d^2 \leq 0.$$

Now we use the mediant $(a + c)/(b + d)$; either ξ is between a/b and the mediant, or ξ is between the mediant and c/d . Let us suppose first that

$$\frac{a}{b} < \xi < \frac{a + c}{b + d}$$

and suppose that the mediant also fails to satisfy the theorem, so that

$$\frac{a + c}{b + d} - \xi \geq \frac{1}{\sqrt{5}(b + d)^2}.$$

Adding corresponding members of this inequality to the first inequality in (7), we can put the result in the form

$$(9) \quad b^2 - \sqrt{5}b(b + d) + (b + d)^2 \leq 0.$$

We will show that (8) and (9) cannot be satisfied simultaneously by $b, d \in \mathcal{L}^+$.

Divide both sides of (8) by b^2 and then let $u = d/b$ to get

$$(10) \quad u^2 - \sqrt{5}u + 1 \leq 0, \quad u = \frac{d}{b}.$$

Divide (9) by b^2 and let $u = 1 + d/b$ to get

$$(11) \quad u^2 - \sqrt{5}u + 1 \leq 0, \quad u = 1 + \frac{d}{b}.$$

Since $f(u) = u^2 - \sqrt{5}u + 1$ is a parabola which opens up, it is clear that $f(u) \leq 0$ if and only if u is between the u -intercepts of the graph of f , hence

$$f(u) \leq 0 \quad \text{if and only if} \quad \frac{\sqrt{5} - 1}{2} \leq u \leq \frac{\sqrt{5} + 1}{2}.$$

Consequently, (10) and (11) are satisfied if and only if

$$\frac{\sqrt{5} - 1}{2} \leq \frac{d}{b} \leq \frac{\sqrt{5} + 1}{2} \quad \text{and} \quad \frac{\sqrt{5} - 1}{2} \leq 1 + \frac{d}{b} \leq \frac{\sqrt{5} + 1}{2}.$$

These conditions are both satisfied only for $d/b = (\sqrt{5} - 1)/2$, which is impossible because d/b is rational and $(\sqrt{5} - 1)/2$ is irrational. Therefore, in the case when ξ is between a/b and the mediant, at least one of our assumptions is untenable, so at least one of a/b , $(a + c)/(b + d)$, b/d satisfies the inequality of the theorem.

In case ξ is between the mediant and c/d , we again assume the mediant fails to satisfy the theorem and we have

$$\xi - \frac{a+c}{b+d} \geq \frac{1}{\sqrt{5}(b+d)^2}.$$

Adding this to the second inequality in (7), we put the result in the form

$$(12) \quad d^2 - \sqrt{5}d(b+d) + (b+d)^2 \leq 0.$$

Upon dividing through by d^2 in (8) and (12), we get $u^2 - \sqrt{5}u + 1 \leq 0$, with $u = b/d$ and $u = 1 + b/d$, respectively. Arguing as in the previous case, we find that $b/d = (\sqrt{5} - 1)/2$, again a contradiction.

We have now proved that if $a/b < c/d$ are successive in $\mathcal{F}_n$ and if $a/b < \xi < c/d$, then one of the numbers $a/b, (a+c)/(b+d), c/d$ gives a solution x/y of the inequality stated in the theorem, and we can inductively describe a construction which provides infinitely many solutions. Suppose for definiteness that

$$\frac{a+c}{b+d} < \xi < \frac{c}{d}.$$

In Exercise 33-4, we take $x' = (a+c)/(b+d)$ and $x'' = \xi$ to find a Farey sequence $\mathcal{F}_{m'}$ containing a'/b' between x' and x'' ; using Exercise 33-4 again with $x' = \xi$ and $x'' = c/d$ we find $\mathcal{F}_{m''}$ containing c'/d' between x' and x'' . If $m = \max(m', m'')$, then $\mathcal{F}_m$ contains both a'/b' and c'/d' ; by relabeling the fractions, if necessary, we can take a'/b' to be the largest fraction in $\mathcal{F}_m$ which does not exceed ξ , and c'/d' to be the smallest in $\mathcal{F}_m$ which ξ does not exceed. Hence $a'/b' < c'/d'$ are successive and ξ is between them; therefore one of $a'/b', (a'+c')/(b'+d'), c'/d'$ gives a solution of the inequality, and this new solution is necessarily different from the previous solution.

We will now prove that all rational numbers which are close (in the sense of Theorem 35.4) to an irrational ξ are convergents to ξ . We require the following.

Lemma 35.4. *Suppose ξ is irrational and P_k/Q_k is the k^{th} convergent to ξ . If $a \in \mathcal{L}$, $b \in \mathcal{L}^+$, and for some $n \geq 0$, $b < Q_{n+1}$, then $|b\xi - a| \geq |Q_n\xi - P_n|$.*

Proof. Consider the system of equations

$$(13) \quad \begin{cases} xP_n + yP_{n+1} = a \\ xQ_n + yQ_{n+1} = b. \end{cases}$$

The unique solution is $x = \Delta(n)(aQ_{n+1} - bP_{n+1}) = \pm(aQ_{n+1} - bP_{n+1})$, $y = \pm(bP_n - aQ_n)$; here we have used Theorem 31.3. Notice that $x, y \in \mathcal{L}$.

Clearly $x \neq 0$, because if $x = 0$, then $aQ_{n+1} = bP_{n+1}$, and $(Q_{n+1}, P_{n+1}) = 1$ implies $Q_{n+1} | b$, so $Q_{n+1} \leq b$, a contradiction. Hence $|x| \geq 1$; if $y = 0$, the proof is complete because $y = 0$ implies $xQ_n = b$ and $xP_n = a$ by (13); therefore, $|b\xi - a| = |x||Q_n\xi - P_n| \geq |Q_n\xi - P_n|$. If $y \neq 0$, then $xy < 0$, because if $x < 0$, then $0 > xQ_n = b - yQ_{n+1}$, so $y > b/Q_{n+1} > 0$; and if $x > 0$, then $yQ_{n+1} = b - xQ_n < b < Q_{n+1}$, so $y < 0$.

Thus, if $y \neq 0$, x and y have opposite signs. Also $Q_{n+1}\xi - P_{n+1}$ and $Q_n\xi - P_n$ have opposite signs because ξ is between the two consecutive convergents P_{n+1}/Q_{n+1} and P_n/Q_n . Therefore the numbers $x(Q_n\xi - P_n)$ and $y(Q_{n+1}\xi - P_{n+1})$ have the same sign. Consequently, the absolute value of their sum is the sum of their absolute values, and

$$\begin{aligned} |b\xi - a| &= |(xQ_n + yQ_{n+1})\xi - (xP_{n+1} + yP_{n+1})| \\ &= |x(Q_n\xi - P_n) + y(Q_{n+1}\xi - P_{n+1})| \\ &= |x||Q_n\xi - P_n| + |y||Q_{n+1}\xi - P_{n+1}| \\ &> |Q_n\xi - P_n|. \end{aligned}$$

Theorem 35.4. *If ξ is irrational and if there exist $a, b \in \mathcal{Z}$, $b \geq 1$, such that*

$$\left| \xi - \frac{a}{b} \right| < \frac{1}{2b^2},$$

then a/b is a convergent to ξ .

Proof. We may assume $(a, b) = 1$ and suppose a/b is not a convergent to ξ . Then we can find the unique n such that $Q_n \leq b < Q_{n+1}$. For this n , $|Q_n\xi - P_n| \leq |b\xi - a| < 1/2b$, so

$$\left| \xi - \frac{P_n}{Q_n} \right| < \frac{1}{2bQ_n}.$$

Since $a/b \neq P_n/Q_n$, we have $bP_n - aQ_n$ is a non-zero integer, and

$$\frac{1}{bQ_n} \leq \frac{|bP_n - aQ_n|}{bQ_n} = \left| \frac{P_n}{Q_n} - \frac{a}{b} \right| \leq \left| \frac{P_n}{Q_n} - \xi \right| + \left| \xi - \frac{a}{b} \right| < \frac{1}{2bQ_n} + \frac{1}{2b^2},$$

this leads to the contradiction $b < Q_n$, and the proof is complete.

Corollary 35.4a. *Every solution x/y of the inequality in the Hurwitz Theorem is a convergent to ξ .*

Corollary 35.4b. *If x, y is a positive solution of $x^2 - Dy^2 = 1$, then x/y is a convergent to $\sqrt{D}$.*

Proof. Since

$$|x - \sqrt{D}y| = \frac{1}{|x + \sqrt{D}y|} < \frac{1}{x + y},$$

we have

$$\left| \sqrt{D} - \frac{x}{y} \right| < \frac{1}{y(x+y)} < \frac{1}{2y^2},$$

since

$$x - y > x - \sqrt{D}y > 0.$$

We now prove that the constant $\sqrt{5}$ in the Hurwitz Theorem is best possible.

Theorem 35.5. *If $c > \sqrt{5}$, then there exists an irrational number ξ such that*

$$\left| \xi - \frac{a}{b} \right| < \frac{1}{cb^2}$$

is satisfied by only a finite number of a/b .

Proof. If $c > \sqrt{5} > 2$, then $|\xi - a/b| < 1/cb^2 < 1/2b^2$ can be satisfied only if a/b is a convergent to ξ . Take $\xi = (\sqrt{5} + 1)/2 = \|1, 1, 1, \dots\|$. By Exercise 32-6, $\lim_{n \rightarrow \infty} (\theta_n^{-1} + Q_{n-1}/Q_n) = \sqrt{5}$, so for only finitely many n can we have $\theta_n^{-1} + Q_{n-1}/Q_n > c$. Hence

$$\left| \xi - \frac{P_n}{Q_n} \right| = \left| \frac{\theta_n^{-1}P_n + P_{n-1}}{\theta_n^{-1}Q_n + Q_{n-1}} - \frac{P_n}{Q_n} \right| = \frac{1}{Q_n^2 \left(\theta_n^{-1} + \frac{Q_{n-1}}{Q_n} \right)} < \frac{1}{cQ_n^2}$$

can hold for only a finite number of convergents.

EXERCISES

35-1. In $\mathcal{F}_1$, $(\sqrt{5} - 1)/2$ is between $1/2$ and $1/1$. Find a solution of

$$\left| \frac{\sqrt{5} - 1}{2} - \frac{a}{b} \right| < \frac{1}{\sqrt{5}b^2}.$$

Use this to find a solution of

$$\left| \frac{\sqrt{5} + 5}{2} - \frac{a}{b} \right| < \frac{1}{\sqrt{5}b^2}.$$

35-2. In $\mathcal{F}_6$, $2/5 < 1/2$ are successive and surround $1/\sqrt{6}$. Find a solution of $|1/\sqrt{6} - a/b| < 1/\sqrt{5}b^2$ among the candidates $2/5, 3/7, 1/2$. Now follow the construction defined in the proof of the Hurwitz Theorem to find a solution different from the first.

35-3. Show that $|1/\sqrt{5} - 17/38| < 1/(2(38)^2)$. Now verify that $17/38$ is a convergent to $1/\sqrt{5}$.

Chapter 7

THE EQUATION $x^n + y^n = z^n$, $n \leq 4$

ONE OF THE most widely known conjectures for which no proof has yet been given is “Fermat’s Last Theorem”: The Diophantine equation $x^n + y^n = z^n$ has no solution with $xyz \neq 0$ if $n > 2$. In this chapter we will show that for $n = 2$ the equation has solutions and we will find all of them. Then we will show that for $n = 3$ and $n = 4$ there are no solutions.

36. Pythagorean Triples

If the integers x, y, z satisfy the equation

$$(1) \quad x^2 + y^2 = z^2,$$

then x, y, z is said to be a *Pythagorean triple*. Suppose x, y, z is such a triple and $(x, y, z) = d$. If we put $x = dx_1$, $y = dy_1$, $z = dz_1$, we see that x_1, y_1, z_1 is also a Pythagorean triple and $(x_1, y_1, z_1) = 1$. On the other hand if x, y, z is any solution of (1) and $k \in \mathcal{L}$, then kx, ky, kz is also a solution. Thus, any solution of (1) may be used to find a solution x, y, z such that $(x, y, z) = 1$, and conversely, a solution with $(x, y, z) = 1$ may be used to generate a family of solutions.

A solution x, y, z such that $(x, y, z) = 1$ is called a *primitive* Pythagorean triple. The above remarks indicate that it suffices to find all primitive solutions of (1). We confine our attention to the case $x > 0$, $y > 0$, $z > 0$.

If x, y, z is a primitive positive solution of (1), then $(x, y) = (x, z) = (y, z) = 1$. To see this, suppose $(x, y) = d$. Then $d^2 | (x^2 + y^2) = z^2$, so $d | z$, and $d | (x, y, z)$. The arguments are similar to show $(x, z) = (y, z) = 1$. In particular, then, not both x and y can be even since $(x, y) = 1$. Also, not both can be odd because, if so, x^2 and y^2 are $1 \pmod{4}$, and $2 \equiv x^2 + y^2 \equiv z^2 \pmod{4}$ is impossible. Therefore, x and y must have opposite parity; we assume x is even, y is odd.

Since x is even, write $x = 2x_1$. Then (1) may be written as

$$(2) \quad z^2 - y^2 = (z - y)(z + y) = 4x_1^2.$$

Now since both y, z are odd, $z \pm y$ are even, say $z - y = 2r_1$ and $z + y = 2s_1$. Then (2) becomes

$$(3) \quad r_1 s_1 = x_1^2.$$

Notice that

$$(r_1, s_1) = \left(\frac{z - y}{2}, \frac{z + y}{2} \right) \left| \frac{(z - y) + (z + y)}{2} = z \right.$$

and

$$(r_1, s_1) \left| \frac{(z + y) - (z - y)}{2} = y. \right.$$

Therefore $(r_1, s_1) = 1$, and because of (3), both r_1 and s_1 must be squares, say $r_1 = r^2$ and $s_1 = s^2$. It is easy to see that $z = s^2 + r^2$ and $y = s^2 - r^2$; since y, z are odd we must have s^2, r^2 (and hence s, r) of opposite parity.

Thus, every positive primitive solution of (1) implies the existence of positive integers $r < s$ which are relatively prime, of opposite parity, and such that $x = 2sr$, $y = s^2 - r^2$, and $z = s^2 + r^2$. Conversely, if r, s are as described here, then x, y, z is obviously a positive Pythagorean triple. Moreover, if $d = (x, y, z)$, let $e = (x, y)$. Clearly $d|e$. If $e > 1$, then e has a prime divisor p . Hence $p|(x, y) = (2sr, s^2 - r^2)$, and since $s^2 - r^2$ is odd, we know $p|(sr, s^2 - r^2)$ and $p \neq 2$. Then

$$(4) \quad \begin{aligned} sr &\equiv 0 \pmod{p} \\ s^2 - r^2 &\equiv 0 \pmod{p}. \end{aligned}$$

Since $(s, r) = 1$, we have exactly one of r, s divisible by p , and this is incompatible with (4). Therefore, $d = e = 1$ and x, y, z is primitive.

We have proved Theorem 36.1.

Theorem 36.1. *The positive primitive solutions of $x^2 + y^2 = z^2$ are $x = 2sr$, $y = s^2 - r^2$, $z = s^2 + r^2$, where $0 < r < s$, $(r, s) = 1$, and r, s are of opposite parity.*

EXERCISES

- 36-1. Determine all positive primitive Pythagorean triples x, y, z such that $x < y < z \leq 50$.
- 36-2. How many imprimitive positive Pythagorean triples are there which satisfy $x < y < z \leq 50$?

- 36-3. Find all right triangles with sides of integral lengths and with area 60.
 36-4. If x, y, z is a primitive Pythagorean triple, prove that $x \pm y$ are either $\equiv 1 \pmod{8}$ or $\equiv -1 \pmod{8}$.
 36-5. Suppose x, y, z is a Pythagorean triple, not necessarily primitive. Prove that at least one of x, y, z is divisible by 3, at least one by 4, and at least one by 5. Thus, $60 \mid xyz$.

37. The Equation $x^4 + y^4 = z^4$

We will first prove a more general result than Fermat's Last Theorem for $n = 4$.

Theorem 37.1. *The equation $x^4 + y^4 = z^2$ has no solution in positive integers.*

Proof. Suppose there is a positive triple x, y, z such that $x^4 + y^4 = z^2$. We may assume that $(x, y) = 1$, and hence that $(x^2, y^2, z) = 1$. Further, we may assume that z is the least positive integer for which there exists $x > 0, y > 0$ such that

$$(5) \quad x^4 + y^4 = z^2, \quad (x, y) = 1.$$

Under these conditions, by Theorem 36.1 there are positive integers $r < s$, $(r, s) = 1$, with r and s of opposite parity, and

$$\begin{aligned} x^2 &= 2sr \\ y^2 &= s^2 - r^2 \\ z &= s^2 + r^2. \end{aligned}$$

If s is even, r odd, then $y^2 \equiv -1 \pmod{4}$, which is impossible. Thus, s is odd and r is even, say $r = 2r'$, and we have

$$x^2 = 4sr'.$$

Since $(r', s) = 1$, both r' and s are squares, say $r' = r_1^2$ and $s = s_1^2$. Then from $y^2 + r^2 = s^2$ we get

$$y^2 + (2r_1^2)^2 = (s_1^2)^2.$$

Again by Theorem 36.1, there are integers $0 < r_2 < s_2$, $(r_2, s_2) = 1$, such that

$$\begin{aligned} 2r_1^2 &= 2r_2s_2 \\ y &= s_2^2 - r_2^2 \\ s_1^2 &= s_2^2 + r_2^2. \end{aligned}$$

Since $r_1^2 = r_2 s_2$ and $(r_2, s_2) = 1$, both r_2 and s_2 are squares, say $r_2 = r_3^2$ and $s_2 = s_3^2$. Hence

$$(6) \quad s_1^2 = s_3^4 + r_3^4, \quad (s_3, r_3) = 1.$$

But $0 < s_1 < s_1^2 = s < s^2 + r^2 = z$, so (6) contradicts the way z was chosen in (5).

Corollary 37.1. *The equation $x^4 + y^4 = z^4$ has no solution with $xyz \neq 0$.*

Proof. If this equation had a solution x, y, z such that $xyz \neq 0$, then $|x|, |y|, z^2$ would be a positive solution of the equation in Theorem 37.1.

EXERCISES

37-1. Suppose $m \in \mathscr{L}^+$. Prove the equation $x^{4m} + y^{4m} = z^{4m}$ has no solution in positive integers.

37-2. Show that in order to prove Fermat's Last Theorem, it suffices to prove that $x^p + y^p = z^p$ has no solution with $xyz \neq 0$ for odd primes p .

38. Arithmetic in $\mathscr{K}(\sqrt{-3})$

Before we can prove Fermat's Last Theorem for $n = 3$, we must investigate some properties of the set

$$\mathscr{J} = \left\{ a + b\xi : a, b \in \mathscr{L}, \xi = \frac{-1 + \sqrt{-3}}{2} \right\}$$

of "Jacobian integers" in the imaginary quadratic number field

$$\mathscr{K}(\sqrt{-3}) = \left\{ \frac{a}{c} + \frac{b}{d}\sqrt{-3} : a, b, c, d \in \mathscr{L}, cd \neq 0 \right\}.$$

The discussion will be very similar to the discussion of Gaussian integers. Since $\mathscr{J}$ is contained in the set of complex numbers, two integers in $\mathscr{J}$ are added and multiplied as complex numbers; thus, if $\alpha, \beta \in \mathscr{J}$, then $\alpha \pm \beta, \alpha\beta \in \mathscr{J}$.

Notice that $\xi^2 = -1 - \xi$ and $\xi^3 = 1$. If $\alpha \in \mathscr{J}$, then the *norm* $D(\alpha)$ of α is defined to be $\alpha\bar{\alpha}$, where $\bar{\alpha}$ is the complex conjugate of α . Therefore, if $\alpha = a + b\xi$, then

$$\begin{aligned} D(\alpha) &= D(a + b\xi) = \left(a - \frac{b}{2} + \frac{b\sqrt{-3}}{2} \right) \left(a - \frac{b}{2} - \frac{b\sqrt{-3}}{2} \right) \\ &= a^2 - ab + b^2. \end{aligned}$$

We have $D(\alpha\beta) = D(\alpha)D(\beta)$. An integer $\alpha \in \mathscr{J}$ is a *unit* if and only if $\alpha|1$, hence if and only if $D(\alpha) = 1$. The units in $\mathscr{J}$ are $\pm 1, \pm \xi, \pm \xi^2$, and no others. Two

integers α, β are called *associates* if there is a unit γ such that $\alpha = \beta\gamma$. A non-zero, non-unit δ in $\mathcal{J}$ is called a *prime* if every factorization of δ involves an associate of δ . To describe the primes in $\mathcal{J}$ we need this lemma which gives some information about rational integers n which can be written in the form $n = x^2 + 3y^2$, $x, y \in \mathcal{Z}$.

Lemma 38.1. *If $n \in \mathcal{Z}^+$ and if there is a prime $q \in \mathcal{Z}$ such that $q|n$ and $q \equiv 5 \pmod{6}$, then n cannot be written in the form $n = x^2 + 3y^2$, $(x, y) = 1$, $x, y \in \mathcal{Z}$. Every prime $p \in \mathcal{Z}$, $p \equiv 1 \pmod{6}$ has a representation as $p = x^2 + 3y^2$ with $(x, y) = 1$, $x, y \in \mathcal{Z}$.*

Proof. Suppose there is a prime q such that $q|n$, $q \equiv 5 \pmod{6}$, and $n = a^2 + 3b^2$, $(a, b) = 1$, for some $a, b \in \mathcal{Z}$. Obviously $q \nmid b$, so there exists x such that $bx \equiv 1 \pmod{q}$. Then

$$0 \equiv x^2 n = x^2(a^2 + 3b^2) \equiv (xa)^2 + 3 \pmod{q}.$$

But then $(-3/q) = 1$ and this contradicts the result in Exercise 17-4.

Now suppose $p \in \mathcal{Z}$ is prime and $p \equiv 1 \pmod{6}$. This proof closely parallels the proof that primes congruent to 1 (mod 4) can be written as a sum of squares, and we again use the approximation of Lemma 26.2. We know (from Exercise 17-4) that $(-3/p) = 1$, so there is some u such that $u^2 \equiv -3 \pmod{p}$. Let $z = -u/p$ and $n = [\sqrt{pk}]$, where $k = 1/\sqrt{3}$. There are $a, b \in \mathcal{Z}$ such that

$$\left| -\frac{u}{p} - \frac{a}{b} \right| \leq \frac{1}{b(n+1)} < \frac{1}{b\sqrt{pk}}, \quad 1 \leq b \leq \sqrt{pk}.$$

Take $c = ub + ap$, and find that

$$|c| = \left| -\frac{u}{p} - \frac{a}{b} \right| |pb| < \sqrt{\frac{p}{k}}.$$

Now $0 < c^2 + 3b^2 < p/k + 3pk = 2\sqrt{3}p < 4p$, and $c^2 + 3b^2 \equiv u^2 b^2 + 3b^2 = b^2(u^2 + 3) \equiv 0 \pmod{p}$. Therefore, we have either $c^2 + 3b^2 = p, 2p$, or $3p$.

If $c^2 + 3b^2 = p$ we are through. It is easy to see that $c^2 + 3b^2 = 2p$ is impossible, because if b and c have the same parity then $4|c^2 + 3b^2$, but $4 \nmid 2p$; also, b and c cannot have opposite parity because then $c^2 + 3b^2$ is odd, but $2p$ is even. Finally, if $c^2 + 3b^2 = 3p$, then $3|c$, say $3c_1 = c$, and we find that $3c_1^2 + b^2 = p$. In any case we have $p = x^2 + 3y^2$. If $(x, y) = d$, then $d^2|p$, so $d = 1$.

Theorem 38.1. *The primes in $\mathcal{J}$ are 2, $1 - \xi$, the rational primes $q \equiv 5 \pmod{6}$, and the integers $a + b + 2b\xi$ and $a - b - 2b\xi$, where $a^2 + 3b^2 = p$ is a rational prime congruent to 1 (mod 6), and associates of these.*

Proof. We notice first that if $D(\alpha)$ is prime in $\mathcal{Z}$, then α is prime in $\mathcal{J}$. Also, just as in the case of Gaussian integers, corresponding to each prime α in $\mathcal{J}$, there is one and only one rational prime p such that $\alpha\beta = p$ for some $\beta \in \mathcal{J}$. (The proof of this fact in $\mathcal{J}$, as in $\mathcal{G}$ depends upon the important lemma that if α is prime and $\alpha|\beta\gamma$, then either $\alpha|\beta$ or $\alpha|\gamma$. We assume for now that this is true; an indication of the proof will be given shortly.) We have $p^2 = D(p) = D(\alpha)D(\beta)$. Thus, either $D(\alpha) = p$, or $D(\alpha) = p^2$ and α is an associate of p .

Let $\alpha = a + b\xi$. If $p = 2$, we see that $D(\alpha) = a^2 - ab + b^2 = 2$ is impossible, because a, b must both be even, but then $4|D(\alpha)$ and $4 \nmid 2$. Therefore, the prime α is an associate of 2 in $\mathcal{J}$.

If $p = 3$, then α cannot be an associate of 3, since $3 = (1 - \xi)(2 + \xi)$ is not prime in $\mathcal{J}$. However, $D(1 - \xi) = D(2 + \xi) = 3$ implies that both $1 - \xi$ and $2 + \xi$ are prime in $\mathcal{J}$, and since $-\xi^2(1 - \xi) = 2 + \xi$, they are associates.

If $p \equiv 5 \pmod{6}$, suppose $D(\alpha) = p$. Then $D(\alpha) = D(a + b\xi) = p = a^2 - ab + b^2$. If $(a, b) = d$, then $d^2|p$, so $(a, b) = 1$. Then also $(2a - b, b) = 1$, and $4p = (2a - b)^2 + 3b^2$ contradicts Lemma 38.1. Hence, $D(\alpha) \neq p$, so α is an associate of p .

If $p \equiv 1 \pmod{6}$, we have seen that there are rational integers x, y such that $p = x^2 + 3y^2$, $(x, y) = 1$. Then

$$p = (x + y + 2y\xi)(x - y - 2y\xi).$$

Since $D(x + y + 2y\xi) = D(x - y - 2y\xi) = p$, α is an associate of one of these factors. Finally, the factors of p are not associates, as is easily seen by considering $\pm\xi^k(x + y + 2y\xi)$, $k = 0, 1, 2$.

The proof of unique factorization in $\mathcal{J}$ again follows from a division algorithm. In the proof we use the fact that if $\beta = b + d\xi$, then $\bar{\beta} = b + d\xi^2$, and this holds for all complex numbers, not only for integers in $\mathcal{J}$.

Theorem 38.2. *If $\alpha, \beta \in \mathcal{J}$, $\beta \neq 0$, then there exist $\lambda, \theta \in \mathcal{J}$ such that $\alpha = \beta\lambda + \theta$, $D(\theta) < D(\beta)$.*

Proof. If $\alpha, \beta \in \mathcal{J}$, $\beta \neq 0$, say $\alpha = a + c\xi$ and $\beta = b + d\xi$. Then

$$\begin{aligned} \frac{\alpha}{\beta} &= \frac{(a + c\xi)(b + d\xi^2)}{(b + d\xi)(b + d\xi^2)} = \frac{(ab + cd - ad) + (cb - ad)\xi}{b^2 - bd + d^2} \\ &= x + y\xi, \quad x, y \text{ rational.} \end{aligned}$$

Choose $m, n \in \mathcal{Z}$ such that $|x - m| \leq 1/2$, $|y - n| \leq 1/2$. Take $\lambda = m + n\xi$ and $\theta = \alpha - \beta\lambda$. Then

$$|\theta| = |\beta| \left| \frac{\alpha}{\beta} - \lambda \right| = |\beta| |(x + y\xi) - (m + n\xi)| = |\beta| |(x - m) + (y - n)\xi|$$

$$\begin{aligned}
&= |\beta| \{((x-m) + (y-n)\xi)((x-m) + (y-n)\xi^2)\}^{1/2} \\
&= |\beta| \{(x-m)^2 + (y-n)^2 - (x-m)(y-n)\}^{1/2} \\
&\leq |\beta| (\tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{2} \cdot \tfrac{1}{2})^{1/2} < |\beta|.
\end{aligned}$$

Therefore, $D(\theta) < D(\beta)$.

With the division algorithm we can define a Euclidean algorithm, a gcd which is unique up to associates, and then prove that if a prime $\alpha \in \mathcal{J}$ divides $\beta\lambda$, then either $\alpha|\beta$ or $\alpha|\lambda$. Then the existence and essential uniqueness of factorization of non-zero, non-units into primes in $\mathcal{J}$ can be verified.

EXERCISES

38-1. Show that $\alpha, \beta \in \mathcal{J}$ implies $\alpha \pm \beta, \alpha\beta \in \mathcal{J}$.

38-2. Show that the units in $\mathcal{J}$ are $(-\xi)^k, k = 1, 2, \dots, 6$.

38-3. Prove

(a) $D(\bar{\alpha}) = D(\alpha)$.

(b) $D(n\alpha) = n^2 D(\alpha)$ for all $n \in \mathcal{Z}^+$.

38-4. If $(\alpha, \beta) = \lambda$, show that λ is not unique but is an associate of any other gcd of α, β .

38-5. Factor $408 + 306\xi$ and $6 + 11\xi$ into products of primes in $\mathcal{J}$.

38-6. (See Exercise 28-4.) Prove that the rational prime p

(a) ramifies in $\mathcal{J}$ if and only if $(-3/p)$ is undefined;

(b) stays prime in $\mathcal{J}$ if and only if $(-3/p) = -1$;

(c) splits in $\mathcal{J}$ if and only if $(-3/p) = 1$.

38-7. Prove that if α is prime in $\mathcal{J}$ and $\alpha|\beta\lambda$, then $\alpha|\beta$ or $\alpha|\lambda$.

39. The Equation $x^3 + y^3 = z^3$

To prove that the Fermat equation is impossible for $n = 3$, we will show that the equation $\alpha^3 + \beta^3 = \gamma^3$ is impossible for $\alpha, \beta, \gamma \in \mathcal{J}$ if $\alpha\beta\gamma \neq 0$. The prime $\pi = 1 - \xi$ plays an important role in what follows.

Lemma 39.1. *If $\alpha \in \mathcal{J}$ and $\pi \nmid \alpha$, then either $\pi^4 | (\alpha^3 + 1)$ or $\pi^4 | (\alpha^3 - 1)$.*

Proof. Recall that $\pi\bar{\pi} = 3 = -\xi^2\pi^2$. If $\alpha \in \mathcal{J}$, $\alpha = a + b\xi = a + b - b\pi$. Since $a + b \in \mathcal{Z}$, we find $q, r \in \mathcal{Z}$ such that $a + b = 3q + r$, $r = 0, \pm 1$. Thus,

$$\alpha = r + \pi(-\xi^2\pi q - b),$$

and in general we notice that every $\beta \in \mathcal{J}$ can be written in the form $\beta = \pi\epsilon + r$, with $r = 0, 1$, or -1 . But since $\pi \nmid \alpha$, $r = 0$ is not possible; therefore, with an obvious definition for $\sigma \in \mathcal{J}$, we have $\alpha = \pm 1 + \pi\sigma$. Then

$$\begin{aligned}
\alpha^3 &= \pm 1 + 3\pi\sigma \pm 3\pi^2\sigma^2 + \pi^3\sigma^3 \\
&= \pm 1 \pm (-\xi^2)\pi^4\sigma^2 + \pi^3\sigma(\sigma^2 - \xi^2).
\end{aligned}$$

Evidently, to complete the proof we need only show that $\pi | \sigma(\sigma^2 - \xi^2)$. But this is obvious because if $\sigma = \pi\varepsilon + r$ with $r = 0$, then $\pi | \sigma$, and with $r = \pm 1$, $\pi | (\sigma \mp \xi)$. Therefore, π^4 divides $\alpha^3 + 1$ or $\alpha^3 - 1$.

Theorem 39.1. *If $\alpha^3 + \beta^3 = \gamma^3$, $\alpha, \beta, \gamma \in \mathcal{J}$, then π must divide at least one of α, β, γ .*

Proof. Suppose the contrary. Then by Lemma 39.1 we can write

$$\alpha^3 = \pm 1 + \pi^4 \sigma_1$$

$$\beta^3 = \pm 1 + \pi^4 \sigma_2$$

$$\gamma^3 = \pm 1 + \pi^4 \sigma_3.$$

Therefore, with $\sigma = \sigma_3 - \sigma_1 - \sigma_2$, we have

$$\pm 1 \pm 1 = \pm 1 + \pi^4 \sigma.$$

With the eight possible choices of sign considered, we get either $\pi^4 | 1$ or $\pi^4 | 3$. Neither is possible, since π cannot divide a unit and π^2 is an associate of 3.

Theorem 39.2. *The equation $\gamma^3 = \alpha^3 + \beta^3$ is impossible for $\alpha, \beta, \gamma \in \mathcal{J}$ if $\alpha\beta\gamma \neq 0$.*

Proof. Suppose there are non-zero integers $\alpha, \beta, \gamma \in \mathcal{J}$ such that $\alpha^3 + \beta^3 = \gamma^3$. If α, β, γ have a common factor, we can remove it to obtain a solution in which the integers have no common factor, and then, by familiar arguments, we can conclude that the new integers are relatively prime in pairs. Also, since π divides one of α, β, γ , we may suppose $\pi | \gamma$ [if not, say $\pi | \beta$, we can write $\beta^3 = \gamma^3 + (-\alpha)^3$], and furthermore we may suppose that the power of π in γ is the least positive power which divides any non-zero integer in $\mathcal{J}$ whose cube is a sum of two cubes. Thus we are assuming that

$$(7) \quad \gamma^3 = \alpha^3 + \beta^3, \quad (\alpha, \beta) = (\alpha, \gamma) = (\beta, \gamma) = 1, \quad \pi | \gamma,$$

and γ contains the minimal power of π .

We consider the factorization

$$\gamma^3 = (\alpha + \beta)(\alpha + \beta\xi)(\alpha + \beta\xi^2).$$

Since $\pi | \gamma$, π divides at least one of the factors $\alpha + \beta\xi^k$, $k = 0, 1, 2$. But $(\alpha + \beta) - (\alpha + \beta\xi) = \beta\pi$, and $(\alpha + \beta\xi) - (\alpha + \beta\xi^2) = \beta\xi\pi$, and since π divides one of the factors, it divides all three. Therefore, π divides all of

$$\alpha + \beta\xi, \alpha\xi + \beta = \xi(\alpha + \beta\xi^2), \alpha\xi^2 + \beta\xi^2$$

since these are associates of the original three factors. Also

$$(\alpha + \beta\xi) + (\alpha\xi + \beta) + (\alpha\xi^2 + \beta\xi^2) = (\alpha + \beta)(1 + \xi + \xi^2) = 0.$$

We now take

$$\alpha_1 = (\alpha + \beta\xi)/\pi, \quad \beta_1 = (\alpha\xi + \beta)/\pi, \quad \gamma_1 = (\alpha\xi^2 + \beta\xi^2)/\pi.$$

Suppose $(\alpha_1, \beta_1) = \delta$. Then $\delta(\xi^2\beta_1 - \xi\alpha_1) = \alpha$ and $\delta(\xi^2\alpha_1 - \xi\beta_1) = \beta$, so $\delta = 1$. Since $\alpha_1 + \beta_1 + \gamma_1 = 0$, we also have $(\alpha_1, \gamma_1) = 1$ and $(\beta_1, \gamma_1) = 1$.

From the relation $\alpha_1\beta_1\gamma_1 = (\gamma/\pi)^3$ and the pairwise relative primality of $\alpha_1, \beta_1, \gamma_1$, we can conclude that each is a cube, say

$$\alpha_1 = \alpha_2^3, \quad \beta_1 = \beta_2^3, \quad \gamma_1 = \gamma_2^3.$$

Evidently $(-\gamma_2)^3 = \alpha_2^3 + \beta_2^3$ and one of $\alpha_2, \beta_2, \gamma_2$ is divisible by π . Say $\pi|\gamma_2$. (It is in fact γ_2 which is a multiple of π , but we need not show this.) It remains only to show that the power of π in γ_2 is less than the power in γ . But this is obvious because $\gamma_2|\gamma_1|(\gamma/\pi)$, and we have constructed a triple which contradicts (7).

Corollary 39.2. *The equation $x^3 + y^3 = z^3$ has no solution in rational integers x, y, z such that $xyz \neq 0$.*

Chapter 8

THE PRIME NUMBER THEOREM

40. Introductory Remarks

If the function $\pi(x)$, x real, is defined by

$$\pi(x) = \#\{p: p \text{ is prime, } p \leq x\},$$

then from Exercise 4-7 we see that $\lim_{x \rightarrow \infty} \pi(x) = \infty$. In this chapter we will study the function $\pi(x)$ in some greater detail and we will prove the Prime Number Theorem:

$$\pi(x) \sim \frac{x}{\log x}.$$

This theorem was first proved in 1896 by Hadamard and Vallée de la Poisson independently; both proofs involve the theory of functions of a complex variable. In 1948, Atle Selberg and Paul Erdős, working independently, gave proofs of the Prime Number Theorem which were “elementary” in the sense that the proofs did not use the theory of functions.

In Selberg’s first proof the concepts of limit inferior and limit superior played a basic role. In a second proof, which appeared in 1949, Selberg used nothing deeper than some simple properties of the logarithm function. Both proofs involve the function

$$\theta(x) = \sum_{p \leq x} \log p,$$

the summation extending over all *primes* $p \leq x$; for example, $\theta(9) = \log 2 + \log 3 + \log 5 + \log 7$. A modification of Selberg’s proof was presented in 1951 by Harold Shapiro, and involves instead the function

$$\psi(x) = \sum_{p^a \leq x} \log p,$$

the summation including one term $\log p$ for each power of p which is $\leq x$; for example, $\psi(9) = 3 \log 2 + 2 \log 3 + \log 5 + \log 7$.

It is this latter proof which we will present in this chapter, because this proof depends only upon techniques already introduced and used in Chapter 2 and Chapter 4. We will first prove that $\psi(x) \sim x$, and later we will show this is equivalent to the Prime Number Theorem.

41. Preliminary Results

The notation $f \cdot g$ to denote the Dirichlet convolution product of two arithmetic functions has already been introduced. We now define the natural product fg and sum $f + g$ of two functions of an integral or real variable by

$$\begin{aligned} fg(x) &= f(x)g(x) \\ (f + g)(x) &= f(x) + g(x). \end{aligned}$$

For example, with these notations we would have $\iota_0 \cdot \mu = \varepsilon$, $\iota_0 \mu = \mu$, and $(\iota_0 + \mu)(n) = \mu(n) + 1$ for every $n \in \mathcal{Z}^+$.

Throughout this chapter, we will write

$$\log x = L(x), \quad x > 0.$$

We will use both the notations “log” and “ L ” for this function, depending upon which notation seems more convenient for a particular discussion.

The function $\Lambda(n) \in \mathcal{A}$ is defined at each $n \in \mathcal{Z}^+$ by

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^\beta, \beta \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, of course, p denotes a prime. It is easy to see that $\psi(x)$ is the summatory function of $\Lambda(n)$, that is,

$$\psi(x) = \sum_{n \leq x} \Lambda(n).$$

Lemma 41.1. *The following functional relations are valid:*

- (1) $L = \Lambda \cdot \iota_0$;
- (2) $\mu \cdot L = \Lambda$;
- (3) $(-\mu L) \cdot \iota_0 = \Lambda$;
- (4) $\mu \cdot (LL) = (\Lambda L) + (\Lambda \cdot \Lambda)$.

Proof. If

$$n = \prod_{i=1}^k p_i^{\beta_i},$$

then

$$\begin{aligned} L(n) &= L\left(\prod_{i=1}^k p_i^{\beta_i}\right) = \sum_{i=1}^k L(p_i^{\beta_i}) = \sum_{i=1}^k \beta_i L(p_i) \\ &= \sum_{i=1}^k \sum_{\theta_i=0}^{\beta_i} \Lambda(p_i^{\theta_i}) = \sum_{d|n} \Lambda(d) = \Lambda \cdot \iota_0(n). \end{aligned}$$

This proves (1), and (2) follows by Möbius inversion.

Then

$$\begin{aligned} \Lambda(n) &= \mu \cdot L(n) = \sum_{d|n} \mu(d) L\left(\frac{n}{d}\right) \\ &= \sum_{d|n} \mu(d) \{L(n) - L(d)\} \\ &= L(n) \sum_{d|n} \mu(d) - \sum_{d|n} \mu(d) L(d) \\ &= L(n) \varepsilon(n) - (\mu L) \cdot \iota_0(n); \end{aligned}$$

but $L(n) = 0$ if $n = 1$ and $\varepsilon(n) = 0$ if $n > 1$. Therefore, for all $n \in \mathscr{E}^+$, $\Lambda(n) = (-\mu L) \cdot \iota_0(n)$, which proves (3).

Finally,

$$\begin{aligned} \mu \cdot (LL)(n) &= \sum_{d|n} \mu(d) L\left(\frac{n}{d}\right) L\left(\frac{n}{d}\right) \\ &= \sum_{d|n} \mu(d) L\left(\frac{n}{d}\right) \{L(n) - L(d)\} \\ &= L(n) (\mu \cdot L)(n) - (\mu L) \cdot L(n) \\ &= L(n) \Lambda(n) - (\mu L) \cdot \iota_0 \cdot \Lambda(n) \\ &= (L\Lambda)(n) + (\Lambda \cdot \Lambda)(n) \end{aligned}$$

and this proves (4).

We now would like an asymptotic formula for the summatory function of the right member of (4). In order to get this, we require some lemmas which come as easy consequences of the Euler–McLaurin sum formula (Theorem 20.1) and the inversion technique of Theorem 23.1.

Lemma 41.2.

$$\sum_{n \leq x} LL(n) = \sum_{n \leq x} \log^2 n = x \log^2 x - 2x \log x + 2x + O(\log^2 x)$$

Proof. From Theorem 20.1,

$$\begin{aligned}\sum_{n \leq x} \log^2 n &= \int_1^x \log^2 t \, dt + 2 \int_1^x (t - [t]) \frac{\log t}{t} \, dt + O(\log^2 x) \\ &= (t \log^2 t - 2t \log t + 2t) \Big|_1^x + O\left(\int_1^x \frac{\log t}{t} \, dt\right) + O(\log^2 x) \\ &= x \log^2 x - 2x \log x + 2x + O(1) + O(\log^2 t)_1^x + O(\log^2 x).\end{aligned}$$

Lemma 41.3.

$$\sum_{n \leq x} L(n) = \sum_{n \leq x} \log n = x \log x - x + O(\log x)$$

The proof is left as an exercise.

Lemma 41.4.

$$\sum_{n \leq x} \frac{\log n}{n} = \frac{1}{2} \log^2 x + C_3 + O\left(\frac{\log x}{x}\right)$$

for some constant C_3 .

Proof. Again by Theorem 20.1,

$$\begin{aligned}\sum_{n \leq x} \frac{\log n}{n} &= \int_1^x \frac{\log t}{t} \, dt + \int_1^x \frac{(t - [t])(1 - \log t)}{t^2} \, dt + O\left(\frac{\log x}{x}\right) \\ &= \frac{1}{2} \log^2 t \Big|_1^x + \int_1^\infty \frac{(t - [t])(1 - \log t)}{t^2} \, dt \\ &\quad - \int_x^\infty \frac{(t - [t])(1 - \log t)}{t^2} \, dt + O\left(\frac{\log x}{x}\right)\end{aligned}$$

if the improper integrals converge. But

$$\int_x^\infty \frac{(t - [t])(1 - \log t)}{t^2} \, dt = O\left(\int_x^\infty \frac{1 - \log t}{t^2} \, dt\right) = O\left(\frac{\log t}{t} \Big|_x^\infty\right) = O\left(\frac{\log x}{x}\right),$$

which approaches 0 as $x \rightarrow \infty$. Therefore, the integral from 1 to ∞ converges, say to C_3 , and the proof is complete.

Lemma 41.5.

$$\sum_{n \leq x} \frac{\mu(n)}{n} \log \frac{x}{n} = O(1)$$

Proof. Take $g(x) = x$ in Theorem 23.1. Use Lemma 19.1 and Corollary 23.1. The details are left as an exercise.

Lemma 41.6.

$$\sum_{n \leq x} \frac{\mu(n)}{n} \log^2 \frac{x}{n} = 2 \log x + O(1)$$

Proof. With $g(x) = x \log x$ in Theorem 23.1 we have

$$\begin{aligned} f(x) &= \sum_{n \leq x} g\left(\frac{x}{n}\right) = x \sum_{n \leq x} \frac{1}{n} (\log x - \log n) = x \log x \sum_{n \leq x} \frac{1}{n} - x \sum_{n \leq x} \frac{\log n}{n} \\ &= x \log x \left\{ \log x + \gamma + O\left(\frac{1}{x}\right) \right\} - x \left\{ \frac{1}{2} \log^2 x + C_3 + O\left(\frac{\log x}{x}\right) \right\} \\ &= \frac{1}{2} x \log^2 x + \gamma x \log x - C_3 x + O(\log x). \end{aligned}$$

Therefore, $x \log x = g(x) = \sum_{n \leq x} \mu(n) f\left(\frac{x}{n}\right)$

$$= \sum_{n \leq x} \mu(n) \left\{ \frac{x}{2n} \log^2 \frac{x}{n} + \gamma \frac{x}{n} \log \frac{x}{n} - \frac{C_3 x}{n} + O\left(\log \frac{x}{n}\right) \right\}$$

$$\begin{aligned} &= \frac{x}{2} \sum_{n \leq x} \frac{\mu(n)}{n} \log^2 \frac{x}{n} + \gamma x \sum_{n \leq x} \frac{\mu(n)}{n} \log \frac{x}{n} - C_3 x \sum_{n \leq x} \frac{\mu(n)}{n} + O\left(\sum_{n \leq x} \frac{\log x}{n}\right) \\ &= \frac{x}{2} S_1 + \gamma x O(1) - C_3 x O(1) + O(\log^2 x). \end{aligned}$$

Hence, $S_1 = 2 \log x + O(1) + O\left(\frac{\log^2 x}{x}\right) = 2 \log x + O(1)$.

Theorem 41.1

$$\sum_{n \leq x} ((\Lambda L) + (\Lambda \cdot \Lambda))(n) = 2x \log x + O(x)$$

Proof. By Lemma 41.1(4) and Theorem 22.1,

$$\sum_{n \leq x} ((\Lambda L) + (\Lambda \cdot \Lambda))(n) = \sum_{n \leq x} \mu \cdot (LL)(n) = \sum_{n \leq x} \mu(n) \sum_{d \leq \frac{x}{n}} LL(d)$$

and by Lemma 41.2 this is

$$\begin{aligned} &= \sum_{n \leq x} \mu(n) \left\{ \frac{x}{n} \log^2 \frac{x}{n} - 2 \frac{x}{n} \log \frac{x}{n} + 2 \frac{x}{n} + O\left(\log^2 \frac{x}{n}\right) \right\} \\ &= x \sum_{n \leq x} \frac{\mu(n)}{n} \log^2 \frac{x}{n} - 2x \sum_{n \leq x} \frac{\mu(n)}{n} \log \frac{x}{n} + 2x \sum_{n \leq x} \frac{\mu(n)}{n} + O\left(\sum_{n \leq x} \log^2 \frac{x}{n}\right) \\ &= x \{ 2 \log x + O(1) \} - 2x O(1) + 2x O(1) + O\left(\sum_{n \leq x} \log^2 \frac{x}{n}\right) \\ &= 2x \log x + O(x) + O\left(\sum_{n \leq x} \log^2 \frac{x}{n}\right). \end{aligned}$$

But $\sum \log^2 \frac{x}{n} = \sum (\log^2 x - 2 \log x \log n + \log^2 n)$

$$= [x] \log^2 x - 2 \log x \sum \log n + \sum \log^2 n$$

$$\begin{aligned}
&= \{x + O(1)\} \log^2 x - 2 \log x \{x \log x - x + O(\log x)\} + \\
&\quad x \log^2 x - 2x \log x + 2x + O(\log^2 x) \\
&= x \log^2 x + O(\log^2 x) - 2x \log^2 x + 2x \log x + O(\log^2 x) + \\
&\quad x \log^2 x - 2x \log x + 2x + O(\log^2 x) = 2x + O(\log^2 x) = O(x) .
\end{aligned}$$

EXERCISES

41-1. Prove Lemma 41.3.

41-2. Supply the details for the proof of Lemma 41.5.

41-3.* Prove

$$\sum_{n \leq x} \log \frac{x}{n} = x + O(\log x) .$$

42. The Function $\psi(x)$

We now introduce a technique from the calculus of finite differences which we will employ frequently in what follows. The method is known as "summation by parts" and, like integration by parts, it interchanges the roles of functions under a summation (integration) sign. The technique is concerned with sums of the type

$$(5) \quad f(1)g(1) + \{f(2) - f(1)\}g(2) + \cdots + \{f(m) - f(m-1)\}g(m) .$$

To simplify notation we do not hesitate to write the above sum as

$$(6) \quad \sum_{n \leq m} \{f(n) - f(n-1)\}g(n)$$

even though $f(0)$ may not be zero, or may not even be defined. Whenever we write (5) in the form of (6) it will be understood that we are using the convention that the term $f(0)g(1)$ is zero.

Theorem 42.1. (*Summation by parts.*) Suppose f, g are functions of a real variable, that both are defined for all $x \geq 1$, and that the above convention is observed. Then

$$\sum_{n \leq x} \{f(n) - f(n-1)\}g(n) = f(\lfloor x \rfloor)g(\lfloor x \rfloor) + \sum_{n \leq x-1} \{g(n) - g(n-1)\}f(n)$$

Proof. We have

$$\begin{aligned}
\sum_{n \leq x} \{f(n) - f(n-1)\}g(n) &= \sum_{n \leq x} f(n)g(n) - \sum_{n \leq x} f(n-1)g(n) \\
&= f(\lfloor x \rfloor)g(\lfloor x \rfloor) + \sum_{n \leq x-1} f(n)g(n) - \sum_{2 \leq n \leq x} f(n-1)g(n)
\end{aligned}$$

$$\begin{aligned}
&= f([x])g([x]) + \sum_{n \leq x-1} f(n)g(n) - \sum_{n \leq x-1} f(n)g(n+1) \\
&= f([x])g([x]) + \sum_{n \leq x-1} \{g(n) - g(n+1)\}f(n) .
\end{aligned}$$

Some special cases of this theorem are very familiar. For example, if $g(n) = \iota_0(n) = 1$, then we obtain this well-known result on telescopic series:

$$\sum_{n \leq x} \{f(n) - f(n-1)\} = f([x]) , \quad \text{if } f(0) = 0 .$$

By taking $f(n) = \iota_1(n) = n$, we have

$$(7) \quad \sum_{n \leq x-1} \{g(n) - g(n+1)\}n = -[x]g([x]) + \sum_{n \leq x} g(n) .$$

We have already used this special case of Theorem 42.1 in the proof of Theorem 20.1.

We proceed now to find some important formulas involving the function $\psi(x)$. Our intention is to establish that

$$(8) \quad \psi(x) = O(x) .$$

In order to prove (8) we will first observe that

$$(9) \quad \sum_{n \leq x} \psi\left(\frac{x}{n}\right) = x \log x - x + O(\log x) .$$

Recall that $\sum_{n \leq x} \log n = x \log x - x + O(\log x)$, which was given in Lemma 41.3.

Also $L = \iota_0 \cdot \Lambda$, given in Lemma 41.1(1). Thus,

$$\sum_{n \leq x} \log n = \sum_{n \leq x} \iota_0(n) \sum_{d \leq \frac{x}{n}} \Lambda(d) = \sum_{n \leq x} \psi\left(\frac{x}{n}\right) ,$$

which proves (9).

Now we use the inversion technique of Theorem 23.1 with $g(x) = \psi(x)$. We have

$$f(x) = \sum_{n \leq x} \psi\left(\frac{x}{n}\right) = \sum_{n \leq x} \log n = x \log x - x + O(\log x) ;$$

therefore, $\psi(x) = \sum_{n \leq x} \mu(n) \left\{ \frac{x}{n} \log \frac{x}{n} - \frac{x}{n} + O\left(\log \frac{x}{n}\right) \right\}$

$$= x \sum_{n \leq x} \frac{\mu(n)}{n} \log \frac{x}{n} - x \sum_{n \leq x} \frac{\mu(n)}{n} + O\left(\mu(n) \left| \sum_{n \leq x} \log \frac{x}{n} \right| \right)$$

$$= xO(1) - xO(1) + O(x) = O(x),$$

as required for (8).

Lemma 42.1.

$$\sum_{n \leq x} \Lambda(n) = \psi(x) \log x + O(x)$$

Proof. Since $\psi(n) - \psi(n-1) = \Lambda(n)$, we use Theorem 42.1 to find

$$\begin{aligned} \sum_{n \leq x} \Lambda(n) &= \sum_{n \leq x} \{ \psi(n) - \psi(n-1) \} L(n) \\ &= \psi(x) \log [x] + \sum_{n \leq x-1} \{ \log n - \log (n+1) \} \psi(n) \\ &= \psi(x) \left\{ \log x + O\left(\frac{1}{x}\right) \right\} - \sum_{n \leq x-1} \log \left(1 + \frac{1}{n}\right) \psi(n) \\ &= \psi(x) \log x + O(\psi(x)) O\left(\frac{1}{x}\right) + \sum_{n \leq x-1} O\left(\frac{1}{n}\right) \psi(n) \\ &= \psi(x) \log x + O(1) + O(x). \end{aligned}$$

Lemma 42.2. Let $\Lambda_1(n) = \Lambda(n)/n$. Then

$$\sum_{n \leq x} \Lambda_1(n) = \log x + O(1).$$

Proof. From (9), and its derivation,

$$\begin{aligned} x \log x + O(x) &= \sum_{n \leq x} \Lambda_1(n) = \sum_{n \leq x} \Lambda(n) \sum_{d \leq \frac{x}{n}} 1 = \sum_{n \leq x} \Lambda(n) \left\{ \frac{x}{n} + O(1) \right\} \\ &= x \sum_{n \leq x} \Lambda_1(n) + O\left(\sum_{n \leq x} \Lambda(n) \right) = x \sum_{n \leq x} \Lambda_1(n) + O(x), \end{aligned}$$

from (8). Divide through by x to complete the argument.

Theorem 42.2.

$$\sum_{n \leq x} \Lambda_1 L(n) = \frac{1}{2} \log^2 x + O(\log x)$$

Proof. Let $f(x) = \sum_{n \leq x} \Lambda_1(n) = \log x + O(1)$. Then $\sum_{n \leq x} \Lambda_1 L(n) =$

$$\begin{aligned} \sum_{n \leq x} \{ f(n) - f(n-1) \} L(n) &= f([x]) \log [x] - \sum_{n \leq x-1} f(n) \log \left(1 + \frac{1}{n}\right) \\ &= \left\{ \log x + O\left(\frac{1}{x}\right) + O(1) \right\} \left\{ \log x + O\left(\frac{1}{x}\right) \right\} - \sum_{n \leq x-1} \{ \log n + O(1) \} \log \left(1 + \frac{1}{n}\right) \end{aligned}$$

$$\begin{aligned}
&= \log^2 x + O(\log x) - \sum_{n \leq x-1} (\log n) \log \left(1 + \frac{1}{n}\right) \\
&\quad + O\left(\sum_{n \leq x-1} \log \left(1 + \frac{1}{n}\right)\right) \\
(10) \quad \sum_{n \leq x} \Lambda_1 L(n) &= \log^2 x + O(\log x) - S_2 + O(S_3).
\end{aligned}$$

Consider a McLaurin series expansion of $\log(1+t)$ with $t = 1/n$:

$$\begin{aligned}
\log \left(1 + \frac{1}{n}\right) &= t + \sum_{k=2}^{\infty} \frac{(-1)^{k-1} t^k}{k} = t + O\left(\sum_{k=2}^{\infty} t^k\right) = t + O\left(t^2 \frac{1}{1-t}\right) \\
&= \frac{1}{n} + O(1/n^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_2 &= \sum_{n \leq x-1} (\log n) \left(\frac{1}{n} + O(1/n^2)\right) \\
&= \sum_{n \leq x-1} \frac{\log n}{n} + O\left(\sum_{n \leq x-1} \frac{\log n}{n^2}\right) \\
&= \frac{1}{2} \log^2(x-1) + C_3 + O\left(\frac{\log(x-1)}{x-1}\right) + O\left(\sum_{n=1}^{\infty} \frac{\log n}{n^2}\right) \\
&= \frac{1}{2} \left\{ \log x \left(1 - \frac{1}{x}\right) \right\}^2 + C_3 + O\left(\frac{\log x}{x}\right) + O(1)
\end{aligned}$$

since the infinite series converges. But

$$\begin{aligned}
\left\{ \log x + \log \left(1 - \frac{1}{x}\right) \right\}^2 &= \left\{ \log x + O\left(\frac{1}{x}\right) \right\}^2 \\
&= \log^2 x + O\left(\frac{\log x}{x}\right)
\end{aligned}$$

so that

$$S_2 = \frac{1}{2} \log^2 x + O(1).$$

Also,

$$S_3 = \sum_{n \leq x-1} \log \left(1 + \frac{1}{n}\right) = \sum_{n \leq x-1} \{\log(n+1) - \log n\} = \log[x] = O(\log x).$$

Making appropriate substitutions into (10), we complete the proof.

Corollary 42.2a.

$$\sum_{n \leq x} \Lambda_1(n) \log \frac{x}{n} = \frac{1}{2} \log^2 x + O(\log x)$$

Proof. Since $\log(x/n) = \log x - \log n$,

$$\begin{aligned} \sum_{n \leq x} \Lambda_1(n) \log \frac{x}{n} &= \log x \sum_{n \leq x} \Lambda_1(n) - \sum_{n \leq x} \Lambda_1(n) \log n \\ &= (\log x) \{\log x + O(1)\} - \{\frac{1}{2} \log^2 x + O(\log x)\} \end{aligned}$$

from Theorem 42.2 and Lemma 42.2.

Corollary 42.2b.

$$\sum_{n \leq x} \Lambda_1 \cdot \Lambda_1(n) = \frac{1}{2} \log^2 x + O(\log x)$$

Proof.

$$\begin{aligned} \sum_{n \leq x} \Lambda_1 \cdot \Lambda_1(n) &= \sum_{n \leq x} \Lambda_1(n) \sum_{d \leq x/n} \Lambda_1(d) \\ &= \sum_{n \leq x} \Lambda_1(n) \left\{ \log \frac{x}{n} + O(1) \right\} \\ &= \sum_{n \leq x} \Lambda_1(n) \log \frac{x}{n} + O\left(\sum_{n \leq x} \Lambda_1(n) \right) \\ &= \frac{1}{2} \log^2 x + O(\log x) + O(\log x) \end{aligned}$$

Theorem 42.3.

$$\sum_{n \leq x} \Lambda \cdot (\Lambda L)(n) = \frac{1}{2} \log^2 x \sum_{n \leq x} \Lambda \cdot \Lambda(n) + O(x \log x)$$

Proof. Let

$$S_4 = \sum_{n \leq x} \Lambda \cdot (\Lambda L)(n).$$

Then

$$\begin{aligned} S_4 &= \sum_{n \leq x} \Lambda(n) \sum_{d \leq x/n} \Lambda L(d) \\ &= \sum_{n \leq x} \Lambda(n) \left\{ \psi\left(\frac{x}{n}\right) \log \frac{x}{n} + O\left(\frac{x}{n}\right) \right\}, \quad \text{from Lemma 42.1,} \\ &= \sum_{n \leq x} \Lambda(n) \psi\left(\frac{x}{n}\right) \log \frac{x}{n} + O\left(x \sum_{n \leq x} \Lambda_1(n)\right) \end{aligned}$$

$$\begin{aligned}
&= \log x \sum_{n \leq x} \Lambda(n) \psi\left(\frac{x}{n}\right) - \sum_{n \leq x} \Lambda L(n) \psi\left(\frac{x}{n}\right) + O(x \log x) \\
&= \log x \sum_{n \leq x} \Lambda(n) \sum_{d \leq x/n} \Lambda(d) - \sum_{n \leq x} \Lambda L(n) \sum_{d \leq x/n} \Lambda(d) + O(x \log x) \\
S_4 &= \log x \sum_{n \leq x} \Lambda \cdot \Lambda(n) - S_4 + O(x \log x).
\end{aligned}$$

The proof is completed by solving the last equation for S_4 .

Theorem 42.4.

$$\sum_{n \leq x} \Lambda \cdot (\Lambda L)(n) + \sum_{n \leq x} \Lambda \cdot \Lambda \cdot \Lambda(n) = x \log^2 x + O(x \log x)$$

Proof. From Theorem 41.1 we have

$$\sum_{d \leq x/n} \Lambda L(d) + \sum_{d \leq x/n} \Lambda \cdot \Lambda(d) = \frac{2x}{n} \log \frac{x}{n} + O\left(\frac{x}{n}\right).$$

We multiply each term by $\Lambda(n)$ and then sum over $n \leq x$ to get

$$\begin{aligned}
\sum_{n \leq x} \Lambda(n) \sum_{d \leq x/n} \Lambda L(d) + \sum_{n \leq x} \Lambda(n) \sum_{d \leq x/n} \Lambda \cdot \Lambda(d) &= 2x \sum_{n \leq x} \Lambda_1(n) \log \frac{x}{n} \\
&\quad + O\left(x \sum_{n \leq x} \Lambda_1(n)\right) \\
\sum_{n \leq x} \Lambda \cdot (\Lambda L)(n) + \sum_{n \leq x} \Lambda \cdot \Lambda \cdot \Lambda(n) &= 2x \left\{ \frac{1}{2} \log^2 x + O(\log x) \right\} \\
&\quad + O(x \log x)
\end{aligned}$$

from Corollary 42.2a and Lemma 42.2. The right side of the last equation is obviously $x \log^2 x + O(x \log x)$.

EXERCISES

42-1. Use Theorem 42.3 and 42.4 to prove that

$$\log x \sum_{n \leq x} \Lambda \cdot \Lambda(n) + 2 \sum_{n \leq x} \Lambda \cdot \Lambda \cdot \Lambda(n) = 2x \log^2 x + O(x \log x).$$

42-2.* Use Lemma 42.1 and Theorem 41.1 to prove *Selberg's formula*:

$$(11) \quad \psi(x) \log x + \sum_{n \leq x} \Lambda \cdot \Lambda(n) = 2x \log x + O(x).$$

Hence prove that

$$\psi(x) \log^2 x + \log x \sum_{n \leq x} \Lambda \cdot \Lambda(n) = 2x \log^2 x + O(x \log x).$$

42-3.* Combine the results of the two previous exercises to prove

$$(12) \quad \psi(x) \log^2 x = 2 \sum_{n \leq x} \Lambda \cdot \Lambda \cdot \Lambda(n) + O(x \log x).$$

42-4.* From Exercise 42-2 and Theorem 42.3 prove

$$(13) \quad \psi(x) \log^2 x + 2 \sum_{n \leq x} \Lambda \cdot (\Lambda L)(n) = 2x \log^2 x + O(x \log x).$$

42-5.* Prove

$$\frac{\Lambda \cdot \Lambda(n)}{n} = \Lambda_1 \cdot \Lambda_1(n)$$

for all $n \in \mathcal{L}^+$.

43. A Fundamental Inequality

We now combine the results of the previous section to obtain a basic inequality which will be used to estimate the size of the quotient $\psi(x)/x$. Our work is simplified somewhat if we introduce the function $r(x)$ defined by

$$r(x) = \psi(x) - x.$$

We will also define the functions α, β, γ as follows:

$$\alpha(x) = \sum_{n \leq x} \Lambda L(n),$$

$$\beta(x) = \sum_{n \leq x} \Lambda \cdot \Lambda(n),$$

$$\gamma(x) = (\alpha + \beta)(x).$$

Lemma 43.1.

$$|r(x)| \log^2 x \leq 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x)$$

Proof. In (12) replace $\psi(x)$ by $r(x) + x$. Then

$$\begin{aligned} r(x) \log^2 x + x \log^2 x &= 2 \sum_{n \leq x} (\Lambda \cdot \Lambda) \cdot \Lambda(n) + O(x \log x) \\ &= 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) \sum_{d \leq x/n} \Lambda(d) + O(x \log x) \\ &= 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) \psi\left(\frac{x}{n}\right) + O(x \log x) \\ &= 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) \left\{ r\left(\frac{x}{n}\right) + \frac{x}{n} \right\} + O(x \log x) \end{aligned}$$

$$\begin{aligned}
&= 2x \sum_{n \leq x} \frac{\Lambda \cdot \Lambda(n)}{n} + 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) r\left(\frac{x}{n}\right) + O(x \log x) \\
&= 2x\left\{\frac{1}{2} \log^2 x + O(\log x)\right\} + 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) r\left(\frac{x}{n}\right) \\
&\quad + O(x \log x)
\end{aligned}$$

by Exercise 42.5 and Corollary 42.2b. Therefore,

$$r(x) \log^2 x = 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) r\left(\frac{x}{n}\right) + O(x \log x).$$

Taking absolute values, we have

$$|r(x)| \log^2 x \leq 2 \sum_{n \leq x} \Lambda \cdot \Lambda(n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x)$$

since $\Lambda \cdot \Lambda(n) \geq 0$ for all n .

Lemma 43.2.

$$|r(x)| \log^2 x \leq 2 \sum_{n \leq x} \Lambda L(n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x)$$

Proof. We first notice that

$$\begin{aligned}
\sum_{n \leq x} (\Lambda L) \cdot \Lambda(n) &= \sum_{n \leq x} \Lambda L(n) \sum_{d \leq x/n} \Lambda(d) \\
&= \sum_{n \leq x} \Lambda L(n) \psi\left(\frac{x}{n}\right) \\
&= \sum_{n \leq x} \Lambda L(n) \left\{ r\left(\frac{x}{n}\right) + \frac{x}{n} \right\} \\
&= \sum_{n \leq x} \Lambda L(n) r\left(\frac{x}{n}\right) + x\left\{\frac{1}{2} \log^2 x + O(\log x)\right\}
\end{aligned}$$

by Theorem 42.2. Now in (13) replace $\psi(x)$ by $r(x) + x$ to get

$$\begin{aligned}
r(x) \log^2 x &= -2 \sum_{n \leq x} (\Lambda L) \cdot \Lambda(n) + x \log^2 x + O(x \log x) \\
&= -2 \sum_{n \leq x} \Lambda L(n) \left\{ r\left(\frac{x}{n}\right) + \frac{x}{n} \right\} + x \log^2 x + O(x \log x) \\
&= -2 \sum_{n \leq x} \Lambda L(n) r\left(\frac{x}{n}\right) - 2x\left\{\frac{1}{2} \log^2 x + O(\log x)\right\} \\
&\quad + x \log^2 x + O(x \log x),
\end{aligned}$$

by Theorem 42.2; therefore,

$$r(x) \log^2 x = -2 \sum_{n \leq x} \Lambda L(n) r\left(\frac{x}{n}\right) + O(x \log x).$$

Again the proof is completed by taking absolute values and using the triangle inequality.

Theorem 43.1.

$$|r(x)| \log^2 x \leq 2 \sum_{n \leq x} (\log n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x)$$

Proof. By adding corresponding members of the inequalities of the last two lemmas, we have

$$(14) \quad |r(x)| \log^2 x \leq \sum_{n \leq x} ((\Lambda L) + (\Lambda \cdot \Lambda))(n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x).$$

Since $\Lambda L(n) = \alpha(n) - \alpha(n-1)$ and $\Lambda \cdot \Lambda(n) = \beta(n) - \beta(n-1)$, and $\gamma(n) = 2n \log n + O(n)$ by Theorem 41.1, we get

$$\begin{aligned} & \sum_{n \leq x} ((\Lambda L) + (\Lambda \cdot \Lambda))(n) \left| r\left(\frac{x}{n}\right) \right| \\ &= \sum_{n \leq x} \{ \alpha(n) + \beta(n) - \alpha(n-1) - \beta(n-1) \} \left| r\left(\frac{x}{n}\right) \right| \\ &= \sum_{n \leq x} \{ \gamma(n) - \gamma(n-1) \} \left| r\left(\frac{x}{n}\right) \right| \\ &= \gamma([x]) \left| r\left(\frac{x}{[x]}\right) \right| + \sum_{n \leq x-1} \left\{ \left| r\left(\frac{x}{n}\right) \right| - \left| r\left(\frac{x}{n+1}\right) \right| \right\} \gamma(n) \\ &= \gamma(x) O(1) + \sum_{n \leq x-1} \left\{ \left| r\left(\frac{x}{n}\right) \right| - \left| r\left(\frac{x}{n+1}\right) \right| \right\} \{ 2n \log n + O(n) \} \\ &= O(x \log x) + 2 \sum_{n \leq x-1} n (\log n) \left\{ \left| r\left(\frac{x}{n}\right) \right| - \left| r\left(\frac{x}{n+1}\right) \right| \right\} \\ &\quad + O \left(\sum_{n \leq x-1} n \left\{ \left| r\left(\frac{x}{n}\right) \right| - \left| r\left(\frac{x}{n+1}\right) \right| \right\} \right) \\ (15) \quad &= O(x \log x) + S_5 + O(S_6). \end{aligned}$$

Now we apply summation by parts on S_5 and get

$$S_5 = -2[x] (\log [x]) \left| r\left(\frac{x}{[x]}\right) \right| + 2 \sum_{n \leq x} \{ n \log n - (n-1) \log (n-1) \} \left| r\left(\frac{x}{n}\right) \right|.$$

But

$$[x](\log [x]) \left| r \left(\frac{x}{[x]} \right) \right| = \{x + O(1)\} \left\{ \log x + O \left(\frac{1}{x} \right) \right\} O(1) = O(x \log x)$$

and

$$\begin{aligned} n \log n - (n-1) \log (n-1) &= \log \frac{n^n}{(n-1)^{n-1}} \\ &= \log n \left(\frac{n}{n-1} \right)^{n-1} \\ &= \log n + \log \left(\frac{n}{n-1} \right)^{n-1} \\ &= \log n + O(1) \end{aligned}$$

since $\log (n/(n-1))^{n-1} \rightarrow 1$ as $n \rightarrow \infty$. Therefore,

$$\begin{aligned} S_5 &= O(x \log x) + 2 \sum_{n \leq x} (\log n) \left| r \left(\frac{x}{n} \right) \right| + O \left(\sum_{n \leq x} \left| r \left(\frac{x}{n} \right) \right| \right) \\ &= 2 \sum_{n \leq x} (\log n) \left| r \left(\frac{x}{n} \right) \right| + O(x \log x), \end{aligned}$$

because

$$\sum \left| r \left(\frac{x}{n} \right) \right| \leq \sum \psi \left(\frac{x}{n} \right) + \sum \frac{x}{n} = O(x \log x)$$

by (9).

Furthermore, since $|u| - |v| \leq |u - v|$,

$$\begin{aligned} S_6 &\leq \sum_{n \leq x-1} n \left| r \left(\frac{x}{n} \right) - r \left(\frac{x}{n+1} \right) \right| \\ &= \sum_{n \leq x-1} n \left| \psi \left(\frac{x}{n} \right) - \psi \left(\frac{x}{n+1} \right) - \left(\frac{x}{n} - \frac{x}{n+1} \right) \right| \\ &\leq \sum_{n \leq x-1} n \left\{ \left| \psi \left(\frac{x}{n} \right) - \psi \left(\frac{x}{n+1} \right) \right| + \left| \frac{x}{n} - \frac{x}{n+1} \right| \right\}. \end{aligned}$$

Because

$$\psi \left(\frac{x}{n} \right) - \psi \left(\frac{x}{n+1} \right) \geq 0 \quad \text{and} \quad \frac{x}{n} - \frac{x}{n+1} > 0$$

we may remove the absolute value signs and write

$$\begin{aligned} S_6 &\leq \sum_{n \leq x-1} n \left\{ \psi\left(\frac{x}{n}\right) - \psi\left(\frac{x}{n+1}\right) \right\} + \sum_{n \leq x-1} n \left\{ \frac{x}{n} - \frac{x}{n+1} \right\} \\ &= -[x] \psi\left(\frac{x}{[x]}\right) + \sum_{n \leq x} \psi\left(\frac{x}{n}\right) + \sum_{n \leq x-1} \frac{x}{n+1}, \quad \text{by (7)} \\ &= 0 + O(x \log x) + O(x \log x) \end{aligned}$$

so that $S_6 = O(x \log x)$. Making substitutions for the values of S_5 and S_6 into (15) we have

$$\sum_{n \leq x} ((\Lambda L) + (\Lambda \cdot \Lambda))(n) \left| r\left(\frac{x}{n}\right) \right| = 2 \sum_{n \leq x} (\log n) \left| r\left(\frac{x}{n}\right) \right| + O(x \log x).$$

This into (14) proves the theorem.

44. The Behavior of $\frac{r(x)}{x}$

We will show in this section that $r(x) = o(x)$, or equivalently, that $r(x)/x \rightarrow 0$ as $x \rightarrow \infty$. Our procedure will be first to show that there are infinitely many values n for which $r(n)/n$ is small. Then we will prove that corresponding to each such n , there is an interval (containing n) throughout which the quotient $r(x)/x$ stays small in absolute value. Finally, we will extend these results to show that $r(x) = o(x)$.

Lemma 44.1. *Let $k > 1$ be an arbitrary, fixed real number. There exists a positive constant C_4 (independent of k) and a number x_0 (depending only on k) such that for every $x \geq x_0$ there is an integer $N \in [x, kx]$ for which*

$$\left| \frac{r(N)}{N} \right| < \frac{C_4}{\log k}.$$

Proof. In Lemma 42.2, we proved $\sum \Lambda_1(n) = \log x + O(1)$. But

$$\begin{aligned} \sum_{n \leq x} \Lambda_1(n) &= \sum_{n \leq x} \Lambda(n) \frac{1}{n} = \sum_{n \leq x} \left\{ \psi(n) - \psi(n-1) \right\} \frac{1}{n} \\ &= \psi([x]) \frac{1}{[x]} + \sum_{n \leq x-1} \left\{ \frac{1}{n} - \frac{1}{n+1} \right\} \psi(n) \\ &= \psi(x) O\left(\frac{1}{x}\right) + \sum_{n \leq x-1} \frac{\psi(n)}{n(n+1)} \\ &= O(1) + \sum_{n \leq x-1} \frac{n + r(n)}{n(n+1)} \end{aligned}$$

$$\begin{aligned}
&= 0(1) + \sum_{n \leq x-1} \frac{1}{n+1} + \sum_{n \leq x-1} \frac{r(n)}{n(n+1)} \\
&= 0(1) + \{\log x + 0(1)\} + \sum_{n \leq x-1} \frac{r(n)}{n^2} \left\{ 1 - \frac{1}{n+1} \right\} \\
&= \log x + 0(1) + \sum_{n \leq x-1} \frac{r(n)}{n^2} - \sum_{n \leq x-1} \frac{r(n)}{n^2(n+1)}.
\end{aligned}$$

But $r(n) = \psi(n) - n = 0(n)$, so

$$\sum_{n \leq x-1} \frac{r(n)}{n^2(n+1)} = 0 \left(\sum_{n \leq x-1} \frac{n}{n^2(n+1)} \right) = 0 \left(\sum_{n=1}^{\infty} \frac{1}{n(n+1)} \right) = 0(1)$$

and

$$\sum_{n \leq x-1} \frac{r(n)}{n^2} = \sum_{n \leq x} \frac{r(n)}{n^2} + 0(1).$$

Therefore,

$$\log x + 0(1) + \sum_{n \leq x} \frac{r(n)}{n^2} = \sum_{n \leq x} \Lambda_1(n) = \log x + 0(1)$$

and it follows that

$$\sum_{n \leq x} \frac{r(n)}{n^2} = 0(1).$$

Thus, there is a constant C such that

$$\left| \sum_{n \leq x} \frac{r(n)}{n^2} \right| < C \quad \text{for all } x.$$

Hence, if $\mathcal{J} = (x, kx]$,

$$\sum_{n \in \mathcal{J}} \frac{r(n)}{n^2} = \sum_{n \leq kx} \frac{r(n)}{n^2} - \sum_{n \leq x} \frac{r(n)}{n^2}$$

so

$$\left| \sum_{n \in \mathcal{J}} \frac{r(n)}{n^2} \right| \leq \left| \sum_{n \leq kx} \frac{r(n)}{n^2} \right| + \left| \sum_{n \leq x} \frac{r(n)}{n^2} \right| < 2C.$$

We now argue differently depending on whether $r(n)$ does or does not change sign on the interval $(x, kx]$.

Case 1, $r(n)$ does not change sign for $n \in \mathcal{J}$.

Then

$$\left| \sum_{n \in \mathcal{J}} \frac{r(n)}{n^2} \right| = \sum_{n \in \mathcal{J}} \left| \frac{r(n)}{n^2} \right| < 2C.$$

Let

$$\left| \frac{r(N)}{N} \right| = \min_{n \in \mathcal{J}} \left| \frac{r(n)}{n} \right|.$$

Hence,

$$\begin{aligned} 2C &> \sum_{n \in \mathcal{J}} \left| \frac{r(n)}{n^2} \right| = \sum_{n \in \mathcal{J}} \left| \frac{r(n)}{n} \right| \frac{1}{n} \geq \left| \frac{r(N)}{N} \right| \sum_{n \in \mathcal{J}} \frac{1}{n} \\ &= \left| \frac{r(N)}{N} \right| \left\{ \sum_{n \leq kx} \frac{1}{n} - \sum_{n \leq x} \frac{1}{n} \right\} \\ &= \left| \frac{r(N)}{N} \right| \left\{ \log kx + \gamma + O\left(\frac{1}{x}\right) - \log x - \gamma + O\left(\frac{1}{x}\right) \right\} \\ &= \left| \frac{r(N)}{N} \right| \left\{ \log k + O\left(\frac{1}{x}\right) \right\}. \end{aligned}$$

Thus we have

$$2C > \left| \frac{r(N)}{N} \right| \left\{ \log k + O\left(\frac{1}{x}\right) \right\}.$$

Since $k > 1$, $\frac{1}{2} \log k > 0$; since $O(1/x) \rightarrow 0$ as $x \rightarrow \infty$, there exists x'_0 depending only on k such that if $x \geq x'_0$, then $O(1/x) \geq -\frac{1}{2} \log k$. Therefore, if $x \geq x'_0$, $\log k + O(1/x) \geq \frac{1}{2} \log k > 0$, and

$$\left| \frac{r(N)}{N} \right| < \frac{2C}{\frac{1}{2} \log k} = \frac{4C}{\log k}.$$

In this case, then, we may take $C_4 \geq 4C$.

Case 2, $r(n)$ changes sign on the interval $\mathcal{J}$.

Suppose one sign change occurs at $n = N$. Then

$$\begin{aligned} |r(N)| &< |r(N) - r(N-1)| = |\psi(N) - \psi(N-1) - (N - (N-1))| \\ &\leq \psi(N) - \psi(N-1) + 1 = \Lambda(N) + 1 \leq \log N + 1. \end{aligned}$$

Therefore,

$$\left| \frac{r(N)}{N} \right| < \frac{\log N + 1}{N}.$$

But since $N \in \mathcal{J}$, $x < N$, and the function $(\log y + 1)/y$ is a decreasing function (the derivative is negative) for $y > 1$, we have

$$\frac{\log N + 1}{N} < \frac{\log x + 1}{x}$$

and

$$\left| \frac{r(N)}{N} \right| < \frac{\log x + 1}{x}.$$

Now, $(\log x + 1)/x \rightarrow 0$ as $x \rightarrow \infty$, so there exists x''_0 depending only on k such that if $x \geq x''_0$, then $(\log x + 1)/x < 1/(\log k)$. Hence for $x \geq x''_0$,

$$\left| \frac{r(N)}{N} \right| < \frac{1}{\log k}.$$

and here we could take $C_4 \geq 1$.

We incorporate the results of the two cases into our choices for x_0 and C_4 . For reasons which will become clear in the proof of the next theorem, we also want $C_4 \geq \log^2 4$. Therefore, choose

$$x_0 = \max \{x'_0, x''_0\}$$

$$C_4 = \max \{4C, \log^2 4\}$$

and the proof is complete.

Notice that Lemma 44.1 implies there are infinitely many integers n such that $|r(n)/n|$ is arbitrarily small. For example, we could argue in this way. Let ε be a positive number. Since C_4 does not depend on k , we can take k so large that $C_4/\log k < \varepsilon$, so that when $x \geq x_0$, there is an $N_1 \in [x, kx]$ such that $|r(N_1)/N_1| < \varepsilon$. Now apply the lemma again with $x = N_1 + 1$: there exists $N_2 \in [N_1 + 1, k(N_1 + 1)]$ such that $|r(N_2)/N_2| < \varepsilon$, and obviously $N_2 \neq N_1$. Continue by induction.

Theorem 44.1. *Let δ be any positive number such that $\delta < 8 \log 2$. Then there exists x_0 (depending only on δ) such that if $x \geq x_0$, then the interval $[x, xe^{4C_4/\delta}]$ contains a subinterval $\mathcal{J} = [N, Ne^{\delta/8}]$ such that for all $y \in \mathcal{J}$,*

$$\left| \frac{r(y)}{y} \right| < \delta.$$

Proof. In Lemma 44.1, take $k = \frac{1}{2}e^{4C_4/\delta}$. Then the x_0 of that lemma is completely determined by δ , call it x'_1 ; if $x \geq x'_1$, then there exists $N \in [x, \frac{1}{2}xe^{4C_4/\delta}]$ such that

$$\left| \frac{r(N)}{N} \right| < \frac{C_4}{\log \frac{1}{2}e^{4C_4/\delta}}.$$

But

$$\log \frac{1}{2} e^{4C_4/\delta} = \log \frac{1}{2} + \frac{4C_4}{\delta} = \frac{4C_4}{\delta} - \log 2$$

and $\log 2 \leq 2C_4/\delta$, because if not, then

$$\log 2 > \frac{2C_4}{\delta} \geq \frac{2(\log 4)^2}{8 \log 2} = \log 2.$$

Therefore,

$$\frac{4C_4}{\delta} - \log 2 \geq \frac{2C_4}{\delta}$$

and

$$\left| \frac{r(N)}{N} \right| < \frac{C_4}{4C_4/\delta - \log 2} \leq \frac{C_4}{2C_4/\delta} = \frac{\delta}{2}.$$

We show that this N is the N in the theorem. First, $x \leq N$ is obvious, and

$$N \leq \frac{1}{2} e^{4C_4/\delta} x.$$

But

$$e^{\delta/8} < e^{(8 \log 2)/8} = 2$$

so

$$e^{\delta/8} N < 2N \leq e^{4C_4/\delta} x.$$

This proves that $\mathcal{J} = [N, e^{\delta/8} N] \subset [x, e^{4C_4/\delta} x]$, and it remains only to show that for all $y \in \mathcal{J}$, $|r(y)/y| < \delta$.

Suppose $y \in \mathcal{J}$. In (11), let $x = y$ and then $x = N$, and subtract the two formulas to get

$$\psi(y) \log y - \psi(N) \log N + \sum_{n \in \mathcal{J}_1} \Lambda \cdot \Lambda(n) = 2y \log y - 2N \log N + 0(y) + 0(N)$$

where $\mathcal{J}_1$ is the interval $(N, y]$. Since $y \leq e^{\delta/8} N < 2N$, $0(y) = 0(N)$; since $\Lambda \cdot \Lambda(n) \geq 0$, the above summation is ≥ 0 , and we have

$$\psi(y) \log y - \psi(N) \log N \leq 2y \log y - 2N \log N + 0(N).$$

We rewrite this in the form

$$\{\psi(y) - \psi(N)\} \log y + \psi(N) \log \frac{y}{N} \leq 2(y - N) \log y + 2N \log \frac{y}{N} + 0(N).$$

Now, $y/N < 2$, so $\log(y/N) = 0(1)$, and $\psi(N) = 0(N)$. We use these observations above and get

$$\{\psi(y) - \psi(N)\} \log y \leq 2(y - N) \log y + 0(N).$$

If we divide through by $\log y$, we have the error term $O(N/\log y) = O(N/\log N)$, because $N < y$, so

$$(16) \quad \psi(y) - \psi(N) \leq 2(y - N) + O\left(\frac{N}{\log N}\right).$$

But

$$\begin{aligned} |r(y)| - |r(N)| - (y - N) &\leq |r(y) - r(N)| - (y - N) \\ &\leq |r(y) - r(N) - (y - N)| \\ &= |\psi(y) - \psi(N)| = \psi(y) - \psi(N), \end{aligned}$$

so from (16) we have

$$(17) \quad \begin{aligned} |r(y)| &\leq |r(N)| + 3(y - N) + O\left(\frac{N}{\log N}\right) \\ \left|\frac{r(y)}{y}\right| &\leq \left|\frac{r(N)}{y}\right| + \frac{3(y - N)}{y} + O\left(\frac{N/y}{\log N}\right). \end{aligned}$$

The error term here is $O(1/\log x)$ because $N/y < 1$ and $x \leq N$ implies

$$\frac{N/y}{\log N} < \frac{1}{\log N} \leq \frac{1}{\log x}.$$

Since $O(1/\log x) \rightarrow 0$ as $x \rightarrow \infty$, there exists x'_1 (depending only on δ) such that if $x \geq x'_1$, then $O(1/\log x) < \delta/8$. Choose

$$x_0 = \max \{x'_1, x''_1\}.$$

Then for $x \geq x_0$,

$$\left|\frac{r(N)}{y}\right| = \frac{N}{y} \left|\frac{r(N)}{N}\right| < \frac{\delta}{2}$$

and

$$O\left(\frac{N/y}{\log N}\right) = O\left(\frac{1}{\log x}\right) < \frac{\delta}{8}.$$

Since $y \leq e^{\delta/8}N$, $N/y \geq e^{-\delta/8}$. We would like to show this is $\geq 1 - \delta/8$. This is easy to see because

$$e^{-\delta/8} = \sum_{n=0}^{\infty} \frac{(-\delta/8)^n}{n!} = 1 - \frac{\delta}{8} + \sum_{n=2}^{\infty} \frac{(-\delta/8)^n}{n!} > 1 - \frac{\delta}{8}.$$

Therefore,

$$\frac{3(y - N)}{y} = 3\left(1 - \frac{N}{y}\right) \leq 3\left\{1 - \left(1 - \frac{\delta}{8}\right)\right\} = \frac{3\delta}{8}.$$

Using these estimates in (17), we have: if $x \geq x_0$, then

$$\left| \frac{r(y)}{y} \right| < \frac{\delta}{2} + \frac{3\delta}{8} + \frac{\delta}{8} = \delta \quad \text{for all } y \in \mathcal{J}.$$

The proof is complete.

Theorem 44.2.

$$r(x) = o(x)$$

Proof. We know $r(x) = O(x)$, so there is a constant C_5 such that $|r(x)| < C_5 x$ for all x . Using this in Theorem 43.1, we get

$$\begin{aligned} |r(x)| \log^2 x &\leq 2 \sum_{n \leq x} (\log n) \frac{C_5 x}{n} + O(x \log x) \\ (18) \qquad &= 2C_5 x \sum_{n \leq x} \frac{\log n}{n} + O(x \log x). \end{aligned}$$

Choose a number c such that $0 < c < 1$ and $cC_5 < 8 \log 2$, and then take $\delta = cC_5$. This determines the x_0 of Theorem 44.1. Now in (18), for those terms arising from $x/n \geq x_0$ and in one of the intervals $[N, e^{\delta/8} N]$, we could have gotten a better bound by using $|r(x/n)| < \delta x/n$ instead of $|r(x/n)| < C_5 x/n$, and by so doing we would have improved the estimate by a factor $(C_5 - \delta)$ for those terms where x/n is in one of the intervals.

Let $\theta = e^{4C_4/\delta}$ and $\lambda = e^{\delta/8}$. Then

$$\mathcal{J}_i = [\theta^{i-1}, \theta^i], \quad i = 1, 2, \dots$$

is a sequence of intervals of the type in Theorem 44.1. If we choose a such that $\theta^a \leq x < \theta^{a+1}$, then $a \log \theta \leq \log x < (a+1) \log \theta$, so

$$a = \left[\frac{\log x}{\log \theta} \right] = \left[\frac{\delta \log x}{4C_4} \right] = O(\log x).$$

If $\theta^b \geq x_0$, then each of the intervals $\mathcal{J}_i$ ($i = b+1, \dots, a$) contains a sub-interval $\mathcal{J}_i' = [N_i, \lambda N_i]$ over which $|r(n)/n| < \delta$. We use the improved bound in (18) and get

$$\begin{aligned} |r(x)| \log^2 x &\leq 2C_5 x \sum_{n \leq x} \frac{\log n}{n} - 2(C_5 - \delta)x \sum_{i=b+1}^a \sum_{x/n \in \mathcal{J}_i'} \frac{\log n}{n} + O(x \log x). \\ (19) \end{aligned}$$

To estimate the double summation occurring here, first consider

$$\begin{aligned}
 \sum_{x/n \in \mathcal{J}_i} \frac{\log n}{n} &= \sum_{N_i \leq x/n \leq \lambda N_i} \frac{\log n}{n} \geq \sum_{x/\lambda N_i < n \leq x/N_i} \frac{\log n}{n} \\
 &\geq \sum_{n \leq x/N_i} \frac{\log n}{n} - \sum_{n \leq x/\lambda N_i} \frac{\log n}{n} \\
 &= \left\{ \frac{1}{2} \log^2 \left(\frac{x}{N_i} \right) + C_3 + O \left(\frac{\log (x/N_i)}{x/N_i} \right) \right\} \\
 &\quad - \left\{ \frac{1}{2} \log^2 \left(\frac{x}{\lambda N_i} \right) + C_3 + O \left(\frac{\log (x/\lambda N_i)}{x/\lambda N_i} \right) \right\}
 \end{aligned}$$

by Lemma 41.4. Both of the error terms are $O(1)$, and

$$\begin{aligned}
 \frac{1}{2} \left\{ \left(\log \frac{x}{N} \right)^2 - \left(\log \frac{x}{\lambda N} \right)^2 \right\} &= \frac{1}{2} \left\{ \left(\log \frac{x}{N} \right)^2 - \left(\log \frac{x}{N} - \log \lambda \right)^2 \right\} \\
 &= \frac{1}{2} \left\{ 2(\log \lambda) \log \frac{x}{N} - \log^2 \lambda \right\} = \frac{\delta}{8} \log \frac{x}{N} + O(1).
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \sum_{i=b+1}^a \sum_{x/n \in \mathcal{J}_i} \frac{\log n}{n} &= \sum_{i=b+1}^a \left\{ \frac{\delta}{8} \log \frac{x}{N_i} + O(1) \right\} \\
 &\geq \frac{\delta}{8} \sum_{i=b+1}^{a/2} \log \frac{x}{N_i} + O \left(\sum_{i=b+1}^a 1 \right) \\
 (20) \quad &\geq \frac{\delta}{8} \sum_{i=b+1}^{a/2} \log \frac{x}{\theta^i} + O(a).
 \end{aligned}$$

In this last step, we have used the fact that

$$\mathcal{J}_i = [N, \lambda N_i] \subset \mathcal{J}_i = [\theta^{i-1}, \theta^i]$$

so $x/N_i > x/\theta^i$. Since $i \leq a/2$, $\theta^i \leq \theta^{a/2} \leq x^{1/2}$, so $\log (x/\theta^i) \geq \log x^{1/2} = \frac{1}{2} \log x$. Also, $O(a) = O(\log x)$. In (20), then,

$$\begin{aligned}
 \sum_{i=b+1}^a \sum_{x/n \in \mathcal{J}_i} \frac{\log n}{n} &\geq \frac{\delta}{8} \sum_{i=1}^{a/2} \frac{1}{2} \log x - \frac{\delta}{8} \sum_{i=1}^b \frac{1}{2} \log x + O(\log x) \\
 &= \frac{\delta}{16} (\log x) \frac{a}{2} - \frac{\delta}{16} (\log x) b + O(\log x)
 \end{aligned}$$

$$\begin{aligned}
&= \frac{\delta}{16}(\log x) \left\{ \frac{1}{2} \frac{\delta \log x}{4C_4} + 0(1) \right\} - \frac{\delta}{16}(\log x) \left[\frac{\log x_0}{\log \theta} \right] \\
&\quad + 0(\log x) \\
&= \frac{\delta^2}{128} \frac{\log^2 x}{C_4} + 0(\log x)
\end{aligned}$$

since $\log x_0 = 0(1)$.

Putting this into (19), and using Lemma 41.4, we have

$$\begin{aligned}
|r(x)| \log^2 x &\leq 2C_5 x \left\{ \frac{1}{2} \log^2 x + C_3 + 0 \left(\frac{\log x}{x} \right) \right\} \\
&\quad - 2(C_5 - \delta)x \left\{ \frac{\delta^2 \log^2 x}{128C_4} + 0(\log x) \right\} + 0(x \log x) \\
&\leq \left\{ C_5 - \frac{(C_5 - \delta)\delta^2}{64C_4} \right\} x \log^2 x + 0(x \log x)
\end{aligned}$$

so that

$$\left| \frac{r(x)}{x} \right| \leq C_5 - \frac{(C_5 - \delta)\delta^2}{64C_4} + 0 \left(\frac{1}{\log x} \right).$$

But $0(1/\log x) \rightarrow 0$ as $x \rightarrow \infty$, so there is some x'_0 such that if $x \geq x'_0$, then

$$0 \left(\frac{1}{\log x} \right) < \frac{(C_5 - \delta)\delta^2}{128C_4}$$

so for $x \geq x_0$ and $x \geq x'_0$, we have

$$\left| \frac{r(x)}{x} \right| \leq C_5 - \frac{(C_5 - \delta)\delta^2}{128C_4}.$$

The bound here is

$$B_1 = C_5 - \frac{(C_5 - \delta)\delta^2}{128C_4} = C_5 - \frac{(C_5 c)^2 C_5 (1 - c)}{128C_4} = C_5 - dC_5^3$$

where

$$d = \frac{(1 - c)c^2}{128C_4} > 0$$

so $B_1 < C_5$. Also, $B_1 > 0$, because

$$\begin{aligned}
128C_4 C_5 - (C_5 - \delta)\delta^2 &\geq 128(\log^2 4)C_5 - (C_5 - \delta)(64 \log^2 2) \\
&= \{128(4 \log^2 2) - 64 \log^2 2\}C_5 + 64\delta \log^2 2 > 0.
\end{aligned}$$

The important thing is that B_1 is a better bound than C_5 . Now we can repeat the process starting with B_1 ; we can use the *same* c (since $cB_1 < cC_5 < 8 \log 2$ allows us to choose $\delta = cB_1$) and this will result in the *same* d (because d depends only on c and C_4 , which is constant), and we will get

$$\left| \frac{r(x)}{x} \right| < B_1 - dB_1^3.$$

Call this $B_2 = B_1 - dB_1^3$, and we have $0 < B_2 < B_1$. Continuing by induction, we find a sequence $B_1 > B_2 > \dots > 0$ which must converge because the sequence is decreasing and bounded. Suppose $B_i \rightarrow B$ as $i \rightarrow \infty$. We have the B_i satisfying $B_{i+1} = B_i - dB_i^3$. Taking limits as $i \rightarrow \infty$, we have $B = B - dB^3$, so $dB^3 = 0$. But $d > 0$, so $B = 0$. Hence, as $x \rightarrow \infty$, $|r(x)/x| \rightarrow 0$, and this is the statement of the theorem.

Corollary 44.2.

$$\psi(x) \sim x$$

Proof. We have $\psi(x) - x = r(x) = o(x)$. Therefore,

$$0 = \lim_{x \rightarrow \infty} \frac{\psi(x) - x}{x} = \lim_{x \rightarrow \infty} \frac{\psi(x)}{x} - 1, \quad \text{or} \quad 1 = \lim_{x \rightarrow \infty} \frac{\psi(x)}{x}.$$

Note. The argument used here applies in general, and the arguments are reversible. Thus $f(x) \sim g(x)$ if and only if $f(x) = g(x) + o(g(x))$.

EXERCISES

44-1.* Prove the statement made in the Note above. Also prove that $f(x) = g(x) + o(g(x))$ if and only if $f(x) = g(x) + o(f(x))$.

45. The Prime Number Theorem and Related Results

The Prime Number Theorem is equivalent to Theorem 44.2. We will first prove that Theorem 44.2 implies the Prime Number Theorem, and then we will give several equivalent statements.

Theorem 45.1. (Prime Number Theorem.)

$$\pi(x) \sim \frac{x}{\log x}$$

Proof. For each prime p , the number of terms $\log p$ in $\psi(x)$ is exactly a , where a satisfies $p^a \leq x < p^{a+1}$. Thus, $a = [(\log x)/(\log p)]$, and

$$\psi(x) = \sum_{p^a \leq x} \log p = \sum_{p \leq x} \left[\frac{\log x}{\log p} \right] \log p \leq \sum_{p \leq x} \log x = \pi(x) \log x$$

so we have

$$\frac{\psi(x)}{x} \leq \frac{\pi(x) \log x}{x}.$$

Now let $b = 1 - 1/\log \log x$. Then $x^b < x$, and

$$\begin{aligned} \pi(x) - x^b &\leq \pi(x) - \pi(x^b) = \sum_{x^b < p \leq x} 1 \leq \sum_{x^b < p \leq x} \frac{\log p}{\log x^b} \\ &= \frac{1}{b \log x} \sum_{x^b < p \leq x} \log p \leq \frac{1}{b \log x} \sum_{p^a \leq x} \log p \\ &= \frac{1}{b \log x} \psi(x). \end{aligned}$$

Therefore,

$$(21) \quad \pi(x) \leq \frac{\psi(x)}{b \log x} + x^b$$

and

$$\frac{\psi(x)}{x} \leq \frac{\pi(x) \log x}{x} \leq \frac{\psi(x)}{bx} + \frac{\log x}{x^{1-b}}.$$

Now take limits as $x \rightarrow \infty$; by Corollary 44.2,

$$1 = \lim_{x \rightarrow \infty} \frac{\psi(x)}{x} \leq \lim_{x \rightarrow \infty} \frac{\pi(x)}{x/\log x} \leq \lim_{x \rightarrow \infty} \frac{\psi(x)}{bx} + \lim_{x \rightarrow \infty} \frac{\log x}{x^{1-b}} = 1 + 0.$$

Hence, $\pi(x) \sim x/\log x$.

Theorem 45.2. *The following are equivalent :*

$$(22) \quad \pi(x) \sim \frac{x}{\log x}$$

$$(23) \quad \psi(x) \sim x$$

$$(24) \quad \theta(x) \sim x.$$

Proof. We have already proved $(23) \Rightarrow (22)$. The reverse implication is proved easily. We get

$$\frac{b\pi(x) \log x}{x} - \frac{b \log x}{x^{1-b}} \leq \frac{\psi(x)}{x} \leq \frac{\pi(x) \log x}{x}$$

where the lower bound is obtained from (21). Now if we assume (22) and notice that $b \rightarrow 1$ as $x \rightarrow \infty$, upon taking limits, (23) follows.

The proof that (24) $\Leftrightarrow$ (22) is almost identical with the proofs already given. Each of the statements in the proofs of Theorems 45.1 and 45.2 can also be made with $\theta(x)$ replacing $\psi(x)$, and the details will be left as an exercise.

Our next theorem, which is also equivalent to the Prime Number Theorem, gives an asymptotic formula for the n^{th} prime.

Theorem 45.3. *If p_n denotes the n^{th} prime, then $p_n \sim n \log n$, and the Prime Number Theorem may be deduced from this relation.*

Proof. First, assume the Prime Number Theorem. Then

$$\frac{p_n}{\log p_n} = \pi(p_n) + o(\pi(p_n)) = n + o(n)$$

and

$$(25) \quad p_n = n \log p_n + o(n \log p_n) = (n \log p_n) \{1 + o(1)\}.$$

Taking logs, we get

$$\log p_n = \log n + \log \log p_n + \log(1 + o(1));$$

but here, $\log \log p_n = o(\log p_n)$, and $\log(1 + o(1)) = o(1)$ since $\lim \log(1 + o(1)) = \log(\lim \{1 + o(1)\}) = \log 1 = 0$. Hence

$$\log p_n = \log n + o(\log p_n),$$

and by Exercise 44-1, $\log n \sim \log p_n$, so $n \log n \sim n \log p_n$. From (25) and Exercise 44-1, we get $n \log p_n \sim p_n$, and by transitivity we get $n \log n \sim p_n$.

Conversely, assume that $p_n \sim n \log n$. This implies (25), and from that, as before, we can derive the relation $\log p_n \sim \log n$. For any x , we find primes p_n, p_{n+1} such that

$$p_n \leq x < p_{n+1}.$$

Hence $\pi(x) = n$. Since the function $f(y) = y/\log y$ is an increasing function, we see that

$$\frac{p_n}{\log p_n} \leq \frac{x}{\log x} < \frac{p_{n+1}}{\log p_{n+1}}.$$

We divide through by $n = \pi(x)$, and get

$$\frac{p_n}{n \log p_n} \leq \frac{x}{\pi(x) \log x} < \frac{p_{n+1}}{n \log p_{n+1}}.$$

Now take limits as $n \rightarrow \infty$ (then also $x \rightarrow \infty$) and we have

$$\frac{p_n}{n \log p_n} \sim \frac{n \log n}{n \log n} = 1 \quad \text{and} \quad \frac{p_{n+1}}{n \log p_{n+1}} \sim \frac{(n+1) \log(n+1)}{n \log(n+1)} \sim 1.$$

Hence,

$$\frac{x}{\pi(x) \log x} \rightarrow 1 \quad \text{as } x \rightarrow \infty,$$

and this is the Prime Number Theorem.

For every $n \in \mathbb{Z}^+$, there exists a prime p such that $n \leq p \leq 2n$. This fact is known as *Bertrand's conjecture*. We will give only an indication of the proof here, but instead we will prove a related result in Theorem 45.4. Since $\theta(x) \sim x$, it is obvious that there are constants C_6 and C_7 such that

$$(26) \quad C_6 x < \theta(x) < C_7 x$$

for x sufficiently large. Since $\theta(x)$ is a step function with a (positive) jump only at the primes, it suffices to prove that $\theta(2n) - \theta(n) > 0$ for all n . But for n sufficiently large, from (26), $\theta(2n) - \theta(n) > (2C_6 - C_7)n$, and by estimating the values of the constants, one proves that $2C_6 - C_7 > 0$; then by direct verification, the result is extended to those n not included in the argument.

Our next theorem is a generalization of Bertrand's conjecture, since the latter is obtained as a special case of Theorem 45.4 by taking $\varepsilon = 1$ and $N = 1$. However, Theorem 45.4 gives no estimate of the value of x_0 , while in the Bertrand conjecture $x_0 = 1$.

Theorem 45.4. *Let ε be any positive real number and N be any positive integer. Then there exists x_0 (depending on ε and on N) such that if $x \geq x_0$, the interval $[x, (1 + \varepsilon)x]$ contains at least N primes.*

Proof. We must show that for x sufficiently large, $\pi((1 + \varepsilon)x) - \pi(x) \geq N$. But

$$(27) \quad \pi((1 + \varepsilon)x) - \pi(x) = \frac{(1 + \varepsilon)x}{\log(1 + \varepsilon)x} - \frac{x}{\log x} + o\left(\frac{(1 + \varepsilon)x}{\log(1 + \varepsilon)x}\right) + o\left(\frac{x}{\log x}\right).$$

We will show that both little- o terms are $o(\varepsilon x / \log(1 + \varepsilon)x)$ and the dominant term in (27) is of the same form.

First, if

$$f(x) = o\left(\frac{(1 + \varepsilon)x}{\log(1 + \varepsilon)x}\right)$$

then

$$0 = \lim_{x \rightarrow \infty} \frac{f(x) \log(1 + \varepsilon)x}{(1 + \varepsilon)x} = \frac{1}{1 + \varepsilon} \lim_{x \rightarrow \infty} \frac{f(x) \log(1 + \varepsilon)x}{x}$$

so that the limit is zero. But then

$$0 = \frac{1}{\varepsilon} \lim_{x \rightarrow \infty} \frac{f(x) \log(1 + \varepsilon)x}{x} = \lim_{x \rightarrow \infty} \frac{f(x)}{\varepsilon x / \log(1 + \varepsilon)x},$$

so

$$f(x) = o\left(\frac{\varepsilon x}{\log(1 + \varepsilon)x}\right).$$

Also

$$0 \leq \frac{|o(x/\log x)|}{\varepsilon x / \log(1 + \varepsilon)x} = \left\{ \frac{1}{\varepsilon} + \frac{\log(1 + \varepsilon)}{\varepsilon \log x} \right\} \frac{|o(x/\log x)|}{x/\log x} \rightarrow 0,$$

so

$$o(x/\log x) = o(\varepsilon x / \log(1 + \varepsilon)x).$$

Finally,

$$\begin{aligned} \frac{(1 + \varepsilon)x}{\log(1 + \varepsilon)x} - \frac{x}{\log x} &= \frac{\varepsilon x}{\log(1 + \varepsilon)x} - \frac{\varepsilon x}{\log(1 + \varepsilon)x} \frac{\log(1 + \varepsilon)}{\varepsilon \log x} \\ &= \frac{\varepsilon x}{\log(1 + \varepsilon)x} + o\left(\frac{\varepsilon x}{\log(1 + \varepsilon)x}\right). \end{aligned}$$

Making substitutions into (27), we conclude

$$\pi((1 + \varepsilon)x) - \pi(x) \sim \frac{\varepsilon x}{\log(1 + \varepsilon)x}.$$

But the function on the right $\rightarrow \infty$, so for x sufficiently large, $\pi((1 + \varepsilon)x) - \pi(x) \geq N$.

Our last application of the Prime Number Theorem is concerned with the function $\text{Li}(x)$, called the “logarithmic integral” function, and defined for all $x > 1$ by

$$\text{Li}(x) = \int_2^x \frac{dt}{\log t}.$$

Theorem 45.5.

$$\text{Li}(x) \sim \pi(x)$$

Proof. We will show that $\text{Li}(x) \sim x/\log x$, and then this theorem follows by a trivial application of Theorem 45.1. Our proof uses the following fact about the Riemann integral of a continuous function: if $f(t)$ is continuous for all t such that $a \leq t \leq b$, and if $f(t)$ is positive for $a \leq t \leq b$, then

$$(b - a) \min_{a \leq t \leq b} f(t) \leq \int_a^b f(t) dt \leq (b - a) \max_{a \leq t \leq b} f(t).$$

It follows from this that such an integral is always non-negative, since $0 \leq (b - a) \min f(t)$.

Evidently $1/\log t$ is positive and decreasing for all $t > 1$. Hence

$$\frac{x-2}{\log x} \leq \text{Li}(x) \leq \frac{x-2}{\log 2}.$$

For our purpose [to show that $\text{Li}(x) \sim x/\log x$], this upper bound is not satisfactory, so we must refine our method.

For each fixed $n \in \mathcal{L}^+$,

$$\begin{aligned} \frac{x}{2^n \log x} \left\{ \frac{1}{1 - \frac{(n-1) \log 2}{\log x}} \right\} &= \frac{x/2^{n-1} - x/2^n}{\log(x/2^{n-1})} \\ &\leq \int_{x/2^n}^{x/2^{n-1}} \frac{dt}{\log t} \\ &\leq \frac{x/2^{n-1} - x/2^n}{\log(x/2^n)} \\ &= \frac{x}{2^n \log x} \left\{ \frac{1}{1 - \frac{n \log 2}{\log x}} \right\}. \end{aligned}$$

Since the integral appearing above is $\text{Li}(x/2^{n-1}) - \text{Li}(x/2^n)$, we get

$$\frac{1}{1 - \frac{(n-1) \log 2}{\log x}} \leq \frac{\text{Li}\left(\frac{x}{2^{n-1}}\right) - \text{Li}\left(\frac{x}{2^n}\right)}{\frac{x}{2^n \log x}} \leq \frac{1}{1 - \frac{n \log 2}{\log x}}.$$

Taking limits as $x \rightarrow \infty$, we prove that

$$\text{Li}\left(\frac{x}{2^{n-1}}\right) - \text{Li}\left(\frac{x}{2^n}\right) \sim \frac{x}{2^n \log x}.$$

By induction on n we will show that we can add corresponding members of the above asymptotic relation and preserve the relation.

$$\begin{aligned} \frac{\text{Li}(x) - \text{Li}(x/2^2)}{\frac{3}{4} \frac{x}{\log x}} &= \frac{\text{Li}(x) - \text{Li}(x/2)}{\frac{3}{4} \frac{x}{\log x}} + \frac{\text{Li}(x/2) - \text{Li}(x/2^2)}{\frac{3}{4} \frac{x}{\log x}} \\ &= \frac{1}{3} \frac{\text{Li}(x) - \text{Li}(x/2)}{\frac{1}{2} \frac{x}{\log x}} + \frac{1}{3} \frac{\text{Li}(x/2) - \text{Li}(x/4)}{\frac{1}{4} \frac{x}{\log x}} \\ &\rightarrow \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 1 = 1. \end{aligned}$$

Therefore, $\text{Li}(x) - \text{Li}(x/4) \sim 3x/4 \log x$. Assume that for $k \geq 2$,

$$(28) \quad \text{Li}(x) - \text{Li}\left(\frac{x}{2^k}\right) \sim \frac{2^k - 1}{2^k} \frac{x}{\log x}.$$

Then

$$\begin{aligned} \frac{\text{Li}(x) - \text{Li}(x/2^{k+1})}{\frac{2^{k+1} - 1}{2^{k+1}} \frac{x}{\log x}} &= \frac{2(2^k - 1)}{2^{k+1} - 1} \frac{\text{Li}(x) - \text{Li}(x/2^k)}{\frac{2^k - 1}{2^k} \frac{x}{\log x}} \\ &\quad + \frac{1}{2^{k+1} - 1} \frac{\text{Li}(x/2^k) - \text{Li}(x/2^{k+1})}{\frac{1}{2^{k+1}} \frac{x}{\log x}} \\ &\rightarrow \frac{2(2^k - 1)}{2^{k+1} - 1} + \frac{1}{2^{k+1} - 1} = 1. \end{aligned}$$

Therefore,

$$\text{Li}(x) - \text{Li}(x/2^{k+1}) \sim \frac{2^{k+1} - 1}{2^{k+1}} \frac{x}{\log x},$$

the induction is complete, and (28) holds for all $k \in \mathcal{Z}^+$.

Let $a + 1 = [\log x / \log 2]$, so $2^{a+1} \leq x < 2^{a+2}$, and $2 \leq x/2^a < 4$. By (28), with $k = a$,

$$\text{Li}(x) - \text{Li}\left(\frac{x}{2^a}\right) \sim \frac{2^a - 1}{2^a} \frac{x}{\log x} \sim \frac{x}{\log x}$$

since as $x \rightarrow \infty$, then $a \rightarrow \infty$, and $(2^a - 1)/2^a \rightarrow 1$. But

$$\text{Li}\left(\frac{x}{2^a}\right) = \int_2^{x/2^a} \frac{dt}{\log t} \leq \int_2^4 \frac{dt}{\log t} = O(1)$$

so that

$$\begin{aligned} 1 &= \lim_{x \rightarrow \infty} \left(\frac{\text{Li}(x)}{x/\log x} - \frac{\text{Li}(x/2^a)}{x/\log x} \right) \\ &= \lim_{x \rightarrow \infty} \frac{\text{Li}(x)}{x/\log x} + \lim_{x \rightarrow \infty} \frac{O(1)}{x/\log x} = \lim_{x \rightarrow \infty} \frac{\text{Li}(x)}{x/\log x} + 0. \end{aligned}$$

This proves $\text{Li}(x) \sim x/\log x$, so $\text{Li}(x) \sim \pi(x)$.

EXERCISES

45-1. Complete the proof of Theorem 45.2 by showing that (24) $\Leftrightarrow$ (23).

45-2. If p_n denotes the n^{th} prime, prove $p_n \sim p_{n+1}$.

45-3. Prove that Theorem 45.5 is equivalent to the Prime Number Theorem.

Chapter 9

GEOMETRY OF NUMBERS

FOR THE MATERIAL in this chapter we assume that the reader has some background in linear algebra. In order to minimize the requirement for a knowledge of linear algebra, we will only study geometry of numbers in the plane; furthermore, our intuition can aid our understanding since we are confining our attention to the 2-dimensional case. Most of the results of this chapter can readily be extended to n -dimensional Euclidean spaces.

The geometry of numbers was first studied by Minkowski. We shall prove some of the fundamental results of Minkowski and apply the results to obtain an upper bound on the least positive value attained by a positive definite binary quadratic form. We will take a second look at Farey sequences from a geometric point of view, and we end our study by using geometric methods to prove that the infinite continued fraction expansion of α is periodic if and only if α is a quadratic irrational.

46. Preliminaries

Let E_2 denote the Cartesian plane. The elements of E_2 will be called *points* or *vectors* and will be denoted by the letters $\mathbf{c}, \mathbf{d}, \mathbf{p}, \mathbf{q}, \dots$. Throughout the chapter the vectors $\mathbf{a}, \mathbf{b}$, and $\mathbf{o}$ will always mean

$$\mathbf{a} = (1, 0),$$

$$\mathbf{b} = (0, 1),$$

$$\mathbf{o} = (0, 0).$$

If $\mathbf{p}, \mathbf{q} \in E_2$, say $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$, then $\mathbf{p} = \mathbf{q}$ if and only if $p_i = q_i$ ($i = 1, 2$), and the vector $\mathbf{p} + \mathbf{q}$, called the *vector sum* of $\mathbf{p}$ and $\mathbf{q}$, is the vector

$$\mathbf{p} + \mathbf{q} = (p_1 + q_1, p_2 + q_2);$$

if x is any real number, then $x\mathbf{p}$ is the vector

$$x\mathbf{p} = (xp_1, xp_2).$$

The vectors $\mathbf{p}$ and $\mathbf{q}$ are said to be *linearly independent* provided that $x\mathbf{p} + y\mathbf{q} = \mathbf{o}$ implies $x = y = 0$. Hence $\mathbf{p}$ and $\mathbf{q}$ are *linearly dependent* if and only if there exists $(x, y) \neq \mathbf{o}$ such that $x\mathbf{p} + y\mathbf{q} = \mathbf{o}$.

If $\mathbf{p}$ and $\mathbf{q}$ are fixed points in $\mathbf{E}_2$, we define the set $\Lambda(\mathbf{p}, \mathbf{q}) \subset \mathbf{E}_2$ by

$$\Lambda(\mathbf{p}, \mathbf{q}) = \{m\mathbf{p} + n\mathbf{q} : m, n \in \mathcal{L}\}.$$

Thus, $\Lambda(\mathbf{p}, \mathbf{q})$ is the set of certain linear combinations of $\mathbf{p}$ and $\mathbf{q}$, and is said to be *generated by* $\mathbf{p}$ and $\mathbf{q}$. The set $\Lambda(\mathbf{p}, \mathbf{q})$ will be called an *integral lattice* if and only if $\mathbf{p}$ and $\mathbf{q}$ are linearly independent, and the points $m\mathbf{p} + n\mathbf{q}$ in an integral lattice are called *lattice points*.

Example. With $\mathbf{p} = \mathbf{a}$ and $\mathbf{q} = \mathbf{b}$ we have

$$\begin{aligned} \Lambda(\mathbf{a}, \mathbf{b}) &= \{m\mathbf{a} + n\mathbf{b} : m, n \in \mathcal{L}\} \\ &= \{(m, n) : m, n \in \mathcal{L}\}. \end{aligned}$$

This integral lattice is just the set of "lattice points" in the sense of Section 17, i.e., the set of all points with integer coordinates. Clearly this Λ is an integral lattice because if $m\mathbf{a} + n\mathbf{b} = (m, n) = \mathbf{o}$, then $m = n = 0$, so $\mathbf{a}$ and $\mathbf{b}$ are linearly independent.

Let $\Lambda_1 = \Lambda(\mathbf{a}, \mathbf{b})$ and $\Lambda_2 = \Lambda((1, 1), (2, 3))$. We notice that $\Lambda_1 = \Lambda_2$: it is obvious that $\Lambda_2 \subset \Lambda_1$ because if $m(1, 1) + n(2, 3) = (m + 2n, m + 3n) \in \Lambda_2$, then the coordinates $m + 2n$ and $m + 3n$ are integers, so the point is in Λ_1 ; conversely, if $(m, n) \in \Lambda_1$, then we find $u, v \in \mathcal{L}$ given by $u = 3m - 2n$, $v = n - m$, and then $(m, n) = u(1, 1) + v(2, 3) \in \Lambda_2$. Here it was possible to find integers u, v as required because there is a solution in integers x, y of the system of equations

$$(1) \quad \begin{cases} x + 2y = m \\ x + 3y = n. \end{cases}$$

Evidently it was fortunate that the determinant of coefficients in (1), namely

$$\begin{vmatrix} 1 & 2 \\ 1 & 3 \end{vmatrix},$$

has the value 1, and we notice that the determinant has for its columns the coordinates of the vectors which generate Λ_2 . These observations shall guide us in introducing certain definitions.

Since the point-pairs $\mathbf{a}, \mathbf{b}$ and $(1, 1), (2, 3)$ generate the same integral lattice, it is natural to regard the point-pairs as related in some sense. We will say

the point-pairs $\mathbf{p}, \mathbf{q}$ and $\mathbf{m}, \mathbf{n}$ are *equivalent*, and we will write $\mathbf{p}, \mathbf{q} \simeq \mathbf{m}, \mathbf{n}$ if and only if $\Lambda(\mathbf{p}, \mathbf{q}) = \Lambda(\mathbf{m}, \mathbf{n})$. It is clear that " $\simeq$ " is an equivalence relation on $\mathbb{E}_2 \times \mathbb{E}_2$, and hence that it partitions the set into equivalence classes. We shall obtain a characterization of these classes.

An integral lattice $\Lambda(\mathbf{p}, \mathbf{q})$ may be viewed geometrically. Consider the vectors $\mathbf{p}$ and $\mathbf{q}$ in the plane, that is, the two directed line segments from $\mathbf{o}$ to $\mathbf{p}$ and from $\mathbf{o}$ to $\mathbf{q}$ (see Figure 6). Let L_0 and L'_0 be the lines containing the

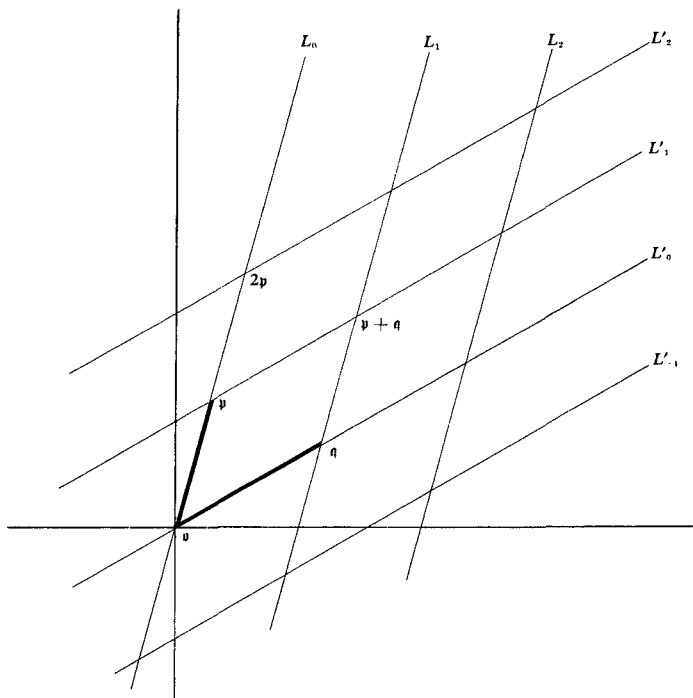


Figure 6

vectors $\mathbf{p}$ and $\mathbf{q}$, respectively. For $j \in \mathcal{Z}$, let L_j be the line parallel to L_0 and through the point $j\mathbf{q}$; let L'_j be the line parallel to L'_0 and through $j\mathbf{p}$. Now the plane is covered by two systems of parallel lines; the collection of points of intersection of these parallel systems is the set of points of $\Lambda(\mathbf{p}, \mathbf{q})$, or

$$\{L_j \cap L'_k : j, k \in \mathcal{Z}\} = \Lambda(\mathbf{p}, \mathbf{q}).$$

If M is any matrix, then $\det M$ will denote the *determinant* of M , $\text{tr } M$ will be the *transpose* of M , and M^{-1} will denote the *inverse* of M if the inverse

exists. We call M an *integral matrix* if the entries of M are integers. If $\mathbf{p}, \mathbf{q} \in \mathbb{E}_2$, we let $(\mathbf{p} \ \mathbf{q})$ denote the matrix whose first column consists of the coordinates of $\mathbf{p}$, and whose second column is $\mathbf{q}$. For example, if $\mathbf{p} = (-\frac{1}{2}, \frac{3}{5})$ and $\mathbf{q} = (2, \frac{5}{4})$, then

$$(\mathbf{p} \ \mathbf{q}) = \begin{pmatrix} -\frac{1}{2} & 2 \\ \frac{3}{5} & \frac{5}{4} \end{pmatrix}$$

and $\det(\mathbf{p} \ \mathbf{q}) = -\frac{73}{40}$.

We can now prove the following theorem.

Theorem 46.1. *Suppose $\mathbf{p}, \mathbf{q}, \mathbf{m}, \mathbf{n} \in \mathbb{E}_2$. Then $\mathbf{p}, \mathbf{q} \simeq \mathbf{m}, \mathbf{n}$ if and only if there exists an integral matrix U with $\det U = \pm 1$ such that $(\mathbf{p} \ \mathbf{q})U = (\mathbf{m} \ \mathbf{n})$.*

Proof. Suppose first that $\mathbf{p}, \mathbf{q} \simeq \mathbf{m}, \mathbf{n}$. Then $\Lambda(\mathbf{p}, \mathbf{q}) = \Lambda(\mathbf{m}, \mathbf{n})$, and in particular $\mathbf{m} \in \Lambda(\mathbf{p}, \mathbf{q})$ and $\mathbf{n} \in \Lambda(\mathbf{p}, \mathbf{q})$. Therefore there are integers a, b, c, d such that $\mathbf{m} = a\mathbf{p} + b\mathbf{q}$ and $\mathbf{n} = c\mathbf{p} + d\mathbf{q}$. Thus, $(\mathbf{p} \ \mathbf{q})U = (\mathbf{m} \ \mathbf{n})$, with

$$U = \begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

an integral matrix. In a similar manner, since $\mathbf{p}, \mathbf{q} \in \Lambda(\mathbf{m}, \mathbf{n})$, there exists an integral matrix V such that $(\mathbf{p} \ \mathbf{q}) = (\mathbf{m} \ \mathbf{n})V$.

Since $\mathbf{p}, \mathbf{q}$ are linearly independent, $\det(\mathbf{p} \ \mathbf{q}) \neq 0$ and $(\mathbf{p} \ \mathbf{q})^{-1}$ exists. Now we have

$$\begin{aligned} UV &= (\mathbf{p} \ \mathbf{q})^{-1}[(\mathbf{p} \ \mathbf{q})U]V \\ &= (\mathbf{p} \ \mathbf{q})^{-1}(\mathbf{m} \ \mathbf{n})V \\ &= (\mathbf{p} \ \mathbf{q})^{-1}(\mathbf{p} \ \mathbf{q}) = I \end{aligned}$$

where I is the 2×2 identity matrix. We know $\det U \in \mathscr{L}$ and $\det V \in \mathscr{L}$ since U and V are integral. But $1 = \det I = \det UV = (\det U)(\det V)$, so $\det U = \pm 1$ as was to be proved. Evidently $V = U^{-1}$ is also integral.

Conversely, assume there is an integral matrix U with determinant ± 1 such that $(\mathbf{p} \ \mathbf{q})U = (\mathbf{m} \ \mathbf{n})$. Suppose $s_1\mathbf{p} + s_2\mathbf{q} \in \Lambda(\mathbf{p}, \mathbf{q})$; if we let $\mathbf{s} = (s_1, s_2)$, we are supposing that $(\mathbf{s})$ is an integral matrix and $(\mathbf{p} \ \mathbf{q})(\mathbf{s}) \in \Lambda(\mathbf{p}, \mathbf{q})$. Clearly, $U^{-1}(\mathbf{s})$ is an integral matrix, say $U^{-1}(\mathbf{s}) = (\mathbf{d})$. But

$$(2) \quad (\mathbf{p} \ \mathbf{q})(\mathbf{s}) = (\mathbf{p} \ \mathbf{q})U(\mathbf{d}) = (\mathbf{m} \ \mathbf{n})(\mathbf{d}),$$

so the point $(\mathbf{p} \ \mathbf{q})(\mathbf{s}) \in \Lambda(\mathbf{m}, \mathbf{n})$, and $\Lambda(\mathbf{p}, \mathbf{q}) \subset \Lambda(\mathbf{m}, \mathbf{n})$. The reverse inclusion is easily proved from the matrix equation (2), and it follows that $\Lambda(\mathbf{m}, \mathbf{n}) = \Lambda(\mathbf{p}, \mathbf{q})$.

EXERCISES

- 46-1. Suppose U is an integral matrix. Find necessary and sufficient conditions so that U^{-1} exists and is integral.
- 46-2. Suppose $\mathbf{p}, \mathbf{q}$ are linearly independent. Prove $\mathbf{p}, \mathbf{q} \simeq \mathbf{q}, \mathbf{p}$.
- 46-3.* If $\mathbf{p}, \mathbf{q} \simeq \mathbf{p}, \mathbf{m}$, prove that there is an integer a such that $\mathbf{q} \pm \mathbf{m} = a\mathbf{p}$.
- 46-4.* In $\Lambda(\mathbf{p}, \mathbf{q})$, consider the parallelogram based on $\mathbf{p}$ and $\mathbf{q}$ (that is, the parallelogram with sides that are the line segments from $\mathbf{o}$ to $\mathbf{p}$, $\mathbf{p}$ to $\mathbf{p} + \mathbf{q}$, $\mathbf{p} + \mathbf{q}$ to $\mathbf{q}$, and $\mathbf{q}$ to $\mathbf{o}$). Prove this parallelogram has area $|\det(\mathbf{p}, \mathbf{q})|$.
- 46-5. Let $\Lambda = \Lambda(\mathbf{p}, \mathbf{q})$. Suppose $\mathbf{m}, \mathbf{n} \in \Lambda$ have the property that $|\det(\mathbf{m}, \mathbf{n})|$ is equal to the area of the parallelogram based on $\mathbf{p}$ and $\mathbf{q}$ (see Exercise 46-4). Prove $\mathbf{m}, \mathbf{n} \simeq \mathbf{p}, \mathbf{q}$.
- 46-6.* Suppose $\mathbf{m}, \mathbf{n} \in \Lambda(\mathbf{p}, \mathbf{q})$. Prove that a necessary and sufficient condition for $\mathbf{m}, \mathbf{n} \simeq \mathbf{p}, \mathbf{q}$ is that the triangle with vertices $\mathbf{o}, \mathbf{m}, \mathbf{n}$ contains no point of Λ except the vertices.
- 46-7. Prove there are lines which contain no points of an integral lattice Λ , and lines which contain exactly one point of Λ , but if a line contains at least two points of Λ then it contains infinitely many.

47. Convex Symmetric Distance Functions

If $\Lambda = \Lambda(\mathbf{p}, \mathbf{q})$ is an integral lattice, we define $\Delta(\Lambda)$ to be $|\det(\mathbf{p}, \mathbf{q})|$. It was shown in Exercise 46-4 that $\Delta(\Lambda)$ is the area of the parallelogram based on $\mathbf{p}$ and $\mathbf{q}$. We denote the *length* of a vector $\mathbf{p} = (p_1, p_2)$ by $|\mathbf{p}|$ and define $|\mathbf{p}| = (p_1^2 + p_2^2)^{1/2}$. Let us assume that $|\mathbf{p}| \leq |\mathbf{q}|$ and consider the circle C with center at the origin and radius $r/2 = (1/2)|\mathbf{p}|$; we would like to show that

$$(3) \quad \pi r^2 < 4\Delta(\Lambda).$$

In Figure 7 we have a circle with radius $r/2$ at each of the vertices of the parallelogram, and it is clear that the sum of the shaded areas is exactly $\pi(r/2)^2$, the area of C . But the shaded areas cannot cover the parallelogram because of the way $r/2$ was chosen, so the area $\Delta(\Lambda)$ is larger than the area of C , which proves (3). Now in turn (3) tells us that the area πr^2 of a circle of radius $r = |\mathbf{p}|$ is less than $4\Delta(\Lambda)$; such a circle is also indicated in Figure 7.

The illustration, which shows how geometric observations can yield arithmetic facts, and vice versa, is a special case of a theorem of Minkowski. Before stating the theorem, we will generalize the important aspects of the above illustration. The circle C will be replaced by any set which is a closed, bounded, convex body symmetric with respect to $\mathbf{o}$ (these terms will be defined below). However, notice that if $f(\mathbf{s}) = |\mathbf{s}|$, then C can be described as the set of points $\mathbf{s}$ such that $f(\mathbf{s}) \leq r/2$; the function f will be replaced by a generalized distance function.

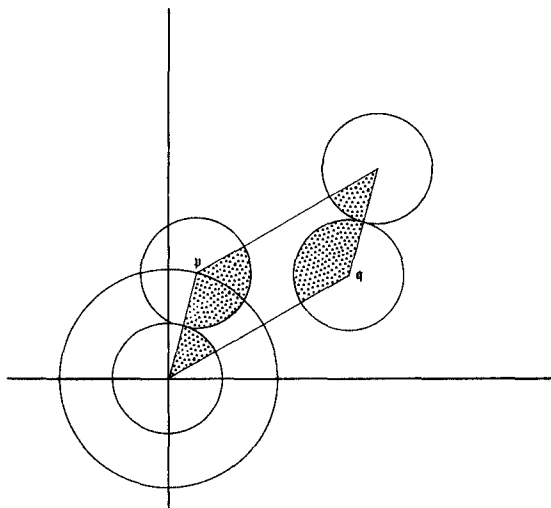


Figure 7

If $\mathcal{S}$ is a set in $\mathbf{E}_2$ and $\mathbf{p} \in \mathcal{S}$, we say $\mathbf{p}$ is an *interior point* of $\mathcal{S}$ if and only if there exists a circle C with center at $\mathbf{p}$ and such that $C \subset \mathcal{S}$. The set $\mathcal{S}$ is *open* if and only if every point of $\mathcal{S}$ is an interior point, and $\mathcal{S}$ is *closed* if and only if its complement (the set of points not in $\mathcal{S}$) is open. The set $\mathcal{S}$ is a *convex set* if and only if the line segment joining any two points of $\mathcal{S}$ lies entirely in $\mathcal{S}$; that is, if $\mathbf{p}, \mathbf{q} \in \mathcal{S}$, then all points

$$(1 - t)\mathbf{p} + t\mathbf{q} \quad (0 \leq t \leq 1)$$

are also in $\mathcal{S}$; a convex set is called a *convex body* if it contains an interior point. The set $\mathcal{S}$ is *bounded* if there is a circle with center at $\mathbf{o}$ which contains $\mathcal{S}$, and $\mathcal{S}$ is said to be *symmetric with respect to $\mathbf{o}$* if $\mathbf{p} \in \mathcal{S}$ implies $-\mathbf{p} \in \mathcal{S}$.

Examples. If $\mathcal{S}$ is a circle, every point of $\mathcal{S}$ not on the circumference is an interior point, and points on the circumference are not interior points. A line has no interior points. A set consisting of a single point has no interior point.

The set of points inside but not on a circle is open; the sets $\mathbf{E}_2$ and $\emptyset$ are open; the set of $\mathbf{s} \neq \mathbf{o}$ is open. The interval $(0, 1)$ is not open. The set of points inside or on a circle is closed; the sets $\mathbf{E}_2$ and $\emptyset$ are closed; a set with a single point is closed; an integral lattice is closed. The interval $(0, 1)$ is not closed.

The set of points inside a circle is convex; the sets $\mathbf{E}_2$ and $\emptyset$ are convex; any line is a convex set; the set of points in a triangle is convex. A set consisting

of a circle and a line is not convex; a set of two points is not convex. The points inside a circle form a convex body; E_2 is a convex body, $\emptyset$ is not a convex body; a line is not a convex body.

An integral lattice is symmetric with respect to $\mathbf{o}$, and so are the sets E_2 and $\emptyset$. A circle with center at $\mathbf{a}$ is not symmetric with respect to $\mathbf{o}$.

Definition. Suppose f is a real-valued function defined for all points $\mathbf{p} \in E_2$. We call f a convex symmetric distance function (csdf) if

- (4) $f(\mathbf{p}) \geq 0$ for all $\mathbf{p} \in E_2$, and $f(\mathbf{p}) = 0$ if and only if $\mathbf{p} = \mathbf{o}$;
- (5) $f(t\mathbf{p}) = |t|f(\mathbf{p})$ for all real t , all $\mathbf{p} \in E_2$;
- (6) $f(\mathbf{p} + \mathbf{q}) \leq f(\mathbf{p}) + f(\mathbf{q})$, for all $\mathbf{p}, \mathbf{q} \in E_2$.

There certainly are such functions because $f(\mathbf{p}) = |\mathbf{p}|$ has the required properties for a csdf. We will show that every csdf is a continuous function. If f is a csdf, then f is continuous at $\mathbf{o}$, because if $\mathbf{p}_1, \mathbf{p}_2, \dots$ is a sequence of points approaching $\mathbf{o}$, say $\mathbf{p}_n = (p_{1n}, p_{2n}) = p_{1n}\mathbf{a} + p_{2n}\mathbf{b}$, then

$$f(\mathbf{p}_n) \leq |p_{1n}|f(\mathbf{a}) + |p_{2n}|f(\mathbf{b}).$$

Let $m = \max(f(\mathbf{a}), f(\mathbf{b}))$; then $0 \leq f(\mathbf{p}_n) \leq m(|p_{1n}| + |p_{2n}|)$ and as $\mathbf{p}_n \rightarrow \mathbf{o}$, $|p_{1n}| + |p_{2n}| \rightarrow 0$, so $f(\mathbf{p}_n) \rightarrow f(\mathbf{o}) = 0$, which proves f is continuous at $\mathbf{o}$. Now if $\mathbf{p}_0 \in E_2$, we prove f is continuous at $\mathbf{p}_0$. Let $\mathbf{p}_n \rightarrow \mathbf{p}_0$, so $\mathbf{p}_n - \mathbf{p}_0 \rightarrow \mathbf{o}$. In (6), we replace $\mathbf{p}$ by $\mathbf{m} - \mathbf{n}$ and $\mathbf{q}$ by $\mathbf{n}$ to see that $f(\mathbf{m} - \mathbf{n}) \geq f(\mathbf{m}) - f(\mathbf{n})$ for all $\mathbf{m}, \mathbf{n} \in E_2$; and in particular,

$$(7) \quad f(\mathbf{p}_n - \mathbf{p}_0) \geq f(\mathbf{p}_n) - f(\mathbf{p}_0).$$

But from $f(\mathbf{p}_n - \mathbf{p}_0) = f(-(\mathbf{p}_0 - \mathbf{p}_n)) = |-1|f(\mathbf{p}_0 - \mathbf{p}_n) \geq f(\mathbf{p}_0) - f(\mathbf{p}_n)$, we have $f(\mathbf{p}_n) - f(\mathbf{p}_0) \geq -f(\mathbf{p}_n - \mathbf{p}_0)$; together with (7) we get

$$-f(\mathbf{p}_n - \mathbf{p}_0) \leq f(\mathbf{p}_n) - f(\mathbf{p}_0) \leq f(\mathbf{p}_n - \mathbf{p}_0).$$

Since $\mathbf{p}_n - \mathbf{p}_0 \rightarrow \mathbf{o}$ and f is continuous at $\mathbf{o}$, we have $f(\mathbf{p}_n) - f(\mathbf{p}_0) \rightarrow 0$, or $f(\mathbf{p}_n) \rightarrow f(\mathbf{p}_0)$. Therefore, f is continuous.

If $k > 0$, we consider the set of all $\mathbf{p}$ such that $f(\mathbf{p}) \leq k$. Since we are interested only in topological properties of the set, we may assume $k = 1$ since $f(\mathbf{p}) \leq k$ if and only if

$$f\left(\frac{1}{k}\mathbf{p}\right) = \frac{1}{k}f(\mathbf{p}) \leq 1.$$

Theorem 47.1. Suppose f is a csdf and let

$$\mathcal{S} = \{\mathbf{p} \in E_2 : f(\mathbf{p}) \leq 1\}.$$

Then $\mathcal{S}$ is a closed, bounded, convex body symmetric with respect to $\mathbf{o}$. A point $\mathbf{p}$ is an interior point of $\mathcal{S}$ if and only if $f(\mathbf{p}) < 1$.

Proof. It is immediate from (5) that $\mathcal{S}$ is symmetric, because $f(-\mathbf{p}) = |-1|f(\mathbf{p})$, so $\mathbf{p} \in \mathcal{S}$ if and only if $-\mathbf{p} \in \mathcal{S}$. The set is convex, because if $\mathbf{p}, \mathbf{q} \in \mathcal{S}$ and t is real, $0 \leq t \leq 1$, then

$$f((1-t)\mathbf{p} + t\mathbf{q}) \leq |1-t|f(\mathbf{p}) + |t|f(\mathbf{q}) \leq |1-t| + |t| = 1;$$

hence $(1-t)\mathbf{p} + t\mathbf{q} \in \mathcal{S}$. But $\mathcal{S}$ also contains interior points, because if $f(\mathbf{p}) < 1$, by the continuity of f , points $\mathbf{q}$ sufficiently close to $\mathbf{p}$ have values $f(\mathbf{q})$ close to $f(\mathbf{p})$. More precisely, there is some $\delta > 0$ such that if $|\mathbf{q} - \mathbf{p}| < \delta$, then $|f(\mathbf{q}) - f(\mathbf{p})| < 1 - f(\mathbf{p})$, or $f(\mathbf{q}) < 1$; but the set of $\mathbf{q}$ such that $|\mathbf{q} - \mathbf{p}| < \delta$ is a circle with center at $\mathbf{p}$, and all such $\mathbf{q}$ satisfy $f(\mathbf{q}) < 1$, so $\mathbf{q} \in \mathcal{S}$. Therefore, $\mathcal{S}$ is a convex body, and we have also proved that if $f(\mathbf{p}) < 1$, then $\mathbf{p}$ is an interior point.

There are points $\mathbf{p}$ which are not in $\mathcal{S}$; for example, since $f(\mathbf{a}) > 0$ and $f(t\mathbf{a}) = |t|f(\mathbf{a})$, we see that the point $t\mathbf{a} \notin \mathcal{S}$ if $t > 1/f(\mathbf{a})$. If $\mathbf{p} \notin \mathcal{S}$, then $f(\mathbf{p}) > 1$ and, again by the continuity of f , if $\mathbf{q}$ is close to $\mathbf{p}$, then $f(\mathbf{q})$ is close to $f(\mathbf{p})$, so that $\mathbf{q} \notin \mathcal{S}$ for all $\mathbf{q}$ sufficiently close to $\mathbf{p}$. Therefore, if $\mathbf{p} \notin \mathcal{S}$, then $\mathbf{p}$ is an interior point of the complement of $\mathcal{S}$, so $\mathcal{S}$ is closed.

To prove boundedness, let Γ denote the unit circle in $\mathbf{E}_2$, that is, $\Gamma = \{\mathbf{p} : \mathbf{p} \in \mathbf{E}_2, |\mathbf{p}| = 1\}$. Clearly Γ is closed and bounded, so f (a continuous function on a closed and bounded set) attains a minimum, say m , on Γ ; hence $m \leq f(\mathbf{p})$ for all $\mathbf{p} \in \Gamma$. If $\mathbf{p} \neq \mathbf{o}$, then

$$\frac{\mathbf{p}}{|\mathbf{p}|} \in \Gamma,$$

so by (5) we have

$$m \leq \frac{1}{|\mathbf{p}|}f(\mathbf{p}) \quad \text{for all } \mathbf{p} \in \mathbf{E}_2,$$

Now $m \neq 0$ because for some $\mathbf{m} \in \Gamma$, we have $f(\mathbf{m}) = m$, and since $\mathbf{m} \in \Gamma$ implies $\mathbf{m} \neq \mathbf{o}$, it follows that $m = f(\mathbf{m}) \neq 0$. Therefore, $|\mathbf{p}| \leq f(\mathbf{p})/m$ for all $\mathbf{p} \in \mathbf{E}_2$; hence, if $\mathbf{p} \in \mathcal{S}$, $|\mathbf{p}| \leq 1/m$ and $\mathcal{S}$ is bounded.

It remains only to show that if $\mathbf{p}$ is an interior point of $\mathcal{S}$, then $f(\mathbf{p}) < 1$. We prove this by showing that if $f(\mathbf{p}) = 1$, then $\mathbf{p}$ is not an interior point of $\mathcal{S}$, because for every $\varepsilon > 0$, $f((1+\varepsilon)\mathbf{p}) = (1+\varepsilon) > 1$, so the points $(1+\varepsilon)\mathbf{p} \notin \mathcal{S}$. But every neighborhood of $\mathbf{p}$ contains the point $(1+\varepsilon)\mathbf{p}$ for ε sufficiently small, so $\mathbf{p}$ is not an interior point of $\mathcal{S}$. This completes the proof of Theorem 47.1.

We will show next that if $\mathcal{K}$ is any closed, bounded, symmetric, convex body, then we can associate with $\mathcal{K}$ a csdf. The details of the proof are a bit

tedious, but the essential idea is simple enough. If $\mathbf{p}$ is any point in E_2 , then the set of all points $t\mathbf{p}$, $0 \leq t$, is a half line from $\mathbf{o}$ and through $\mathbf{p}$. As we move along this line from $\mathbf{o}$ we must come to a point $t_0\mathbf{p}$ which is the last point in $\mathcal{K}$. We will then take $f(t_0\mathbf{p}) = 1$ and $f(r\mathbf{p}) = r/t_0$, so that $f(r\mathbf{p})$ is ≤ 1 or > 1 according as $r \leq t_0$ or $r > t_0$; thus, $f(r\mathbf{p}) \leq 1$ if $r\mathbf{p}$ is between $\mathbf{o}$ and $t_0\mathbf{p}$, otherwise $f(r\mathbf{p}) > 1$.

We first prove the following.

Lemma 47.1. *If $\mathcal{K}$ is a convex body symmetric with respect to $\mathbf{o}$, then $\mathbf{o}$ is an interior point of $\mathcal{K}$.*

Proof. Since $\mathcal{K}$ is a convex body, it contains an interior point $\mathbf{p}$. If $\mathbf{p} = \mathbf{o}$, we are through, so assume $\mathbf{p} \neq \mathbf{o}$. Then $-\mathbf{p} \in \mathcal{K}$, and by convexity $(1 - \frac{1}{2})\mathbf{p} + \frac{1}{2}(-\mathbf{p}) = \mathbf{o} \in \mathcal{K}$. There is a circular disc Γ_1 with center at $\mathbf{p}$ which is in $\mathcal{K}$, say $\Gamma_1 = \{\mathbf{q} : |\mathbf{q} - \mathbf{p}| < \delta\}$. If $\mathbf{q} \in \Gamma_1 \subset \mathcal{K}$, then $-\mathbf{q} \in \mathcal{K}$, so the disc $\Gamma_2 = \{-\mathbf{q} : \mathbf{q} \in \Gamma_1\} = \{\mathbf{q} : |\mathbf{q} + \mathbf{p}| < \delta\} \subset \mathcal{K}$. (See Figure 8.)

The point $\mathbf{s}_1 = \mathbf{p} - (\delta/2)\mathbf{b} \in \Gamma_1$ since $|\mathbf{p} - \mathbf{s}_1| = \delta/2$. Similarly, the point $\mathbf{s}_2 = \mathbf{p} + (\delta/2)\mathbf{b} \in \Gamma_1$; hence $-\mathbf{s}_1, -\mathbf{s}_2 \in \Gamma_2$, so the line segments L_1 and L_2 are in $\mathcal{K}$, where L_1 is from $\mathbf{s}_1$ to $-\mathbf{s}_2$ and L_2 is from $\mathbf{s}_2$ to $-\mathbf{s}_1$. Now consider the region $\mathcal{T}$ (shaded in Figure 8) bounded by the discs Γ_1, Γ_2 and the lines L_1, L_2 ; we show $\mathcal{T} \subset \mathcal{K}$.

If $\mathbf{m} \in \mathcal{T}$, consider a line through $\mathbf{m}$ and parallel to L_1 ; such a line intersects points in Γ_1 and points in Γ_2 , and the segment of such a line from Γ_1 to Γ_2 is

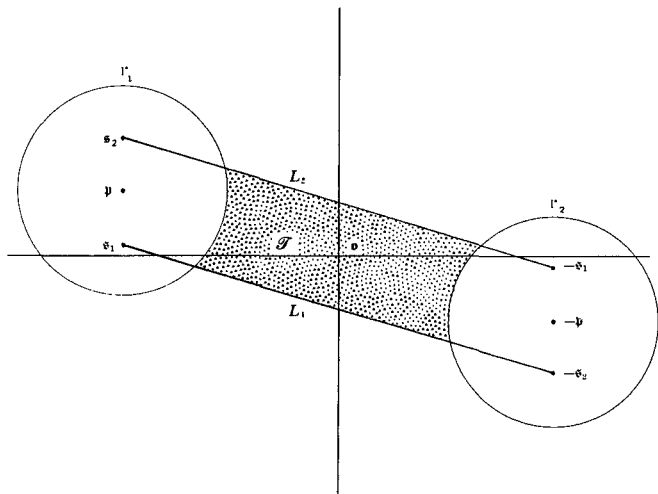


Figure 8

in $\mathcal{K}$ by convexity. Hence $\mathbf{m} \in \mathcal{K}$, so $\mathcal{T} \subset \mathcal{K}$. Evidently we can now find a circle with center at $\mathbf{o}$ and which is in $\mathcal{T}$, hence in $\mathcal{K}$. Therefore, $\mathbf{o}$ is an interior point of $\mathcal{K}$.

Theorem 47.2. Suppose $\mathcal{K}$ is a closed and bounded convex body symmetric with respect to $\mathbf{o}$. Define a function f on $\mathbf{E}_2$ by

$$f(\mathbf{p}) = \begin{matrix} \text{glb } t; \\ t > 0 \\ t^{-1}\mathbf{p} \in \mathcal{K} \end{matrix}$$

then f is a csdf. Furthermore, if $\mathcal{S}$ is the set associated with f as in Theorem 47.1, then $\mathcal{S} = \mathcal{K}$.

Proof. We know from the lemma that there is a circle Γ_3 , center at $\mathbf{o}$, $\Gamma_3 \subset \mathcal{K}$, and since $\mathcal{K}$ is bounded there is a circle Γ_4 such that $\mathcal{K} \subset \Gamma_4$. Suppose the radius of Γ_i is r_i ($i = 3, 4$).

Since $t\mathbf{o} \in \mathcal{K}$ for all $t > 0$, $f(\mathbf{o}) = \text{glb } t = 0$. If $\mathbf{p} \neq \mathbf{o}$, and if $0 < |t^{-1}\mathbf{p}| < r_3$, then $t^{-1}\mathbf{p} \in \Gamma_3 \subset \mathcal{K}$; thus, $t^{-1}\mathbf{p} \in \mathcal{K}$ if $t > |\mathbf{p}|/r_3$, so $f(\mathbf{p}) \leq |\mathbf{p}|/r_3$ is real. Similarly if $t < |\mathbf{p}|/r_4$, then $t^{-1}\mathbf{p} \notin \Gamma_4$, hence $\notin \mathcal{K}$, so $f(\mathbf{p}) \geq |\mathbf{p}|/r_4 > 0$. We have proved that $f(\mathbf{p})$ is real-valued, $f(\mathbf{p}) \geq 0$ for all $\mathbf{p}$, and $f(\mathbf{p}) = 0$ if and only if $\mathbf{p} = \mathbf{o}$.

To prove that $f(k\mathbf{p}) = |k|f(\mathbf{p})$, we use the fact that if $\mathcal{X}, \mathcal{Y}$ are sets of reals and $\mathcal{X} \subset \mathcal{Y}$, then $\text{glb } \mathcal{X} \geq \text{glb } \mathcal{Y}$. Let $\mathcal{A} = \{t: t^{-1}(k\mathbf{p}) \in \mathcal{K}\}$ and $\mathcal{B} = \{r: r^{-1}\mathbf{p} \in \mathcal{K}\}$. If $t \in \mathcal{A}$, then $t|k|^{-1} \in \mathcal{B}$; therefore,

$$|k|^{-1}f(k\mathbf{p}) = |k|^{-1} \begin{matrix} \text{glb } t \\ t \in \mathcal{A} \end{matrix} = \begin{matrix} \text{glb } t|k|^{-1} \\ t \in \mathcal{A} \end{matrix} \geq \begin{matrix} \text{glb } t|k|^{-1} \\ t|k|^{-1} \in \mathcal{B} \end{matrix} = f(\mathbf{p}).$$

Also, if $r \in \mathcal{B}$, then $r|k| \in \mathcal{A}$, so

$$|k|f(\mathbf{p}) = |k| \begin{matrix} \text{glb } r \\ r \in \mathcal{B} \end{matrix} \geq \begin{matrix} \text{glb } r|k| \\ r|k| \in \mathcal{A} \end{matrix} = f(k\mathbf{p}).$$

We now prove that $f(\mathbf{p}) \leq 1$ if and only if $\mathbf{p} \in \mathcal{K}$. If $f(\mathbf{p}) < 1$ and $\mathbf{p} \neq \mathbf{o}$, there is some t such that $0 < f(\mathbf{p}) \leq t < 1$ for which $t^{-1}\mathbf{p} \in \mathcal{K}$. Since $\mathbf{o} \in \mathcal{K}$, convexity implies $(1-t)\mathbf{o} + t(t^{-1}\mathbf{p}) = \mathbf{p} \in \mathcal{K}$. If $f(\mathbf{p}) = 1$, then $\mathbf{p}$ is the limit of a sequence of points $\mathbf{q}$ with $f(\mathbf{q}) < 1$; thus, $\mathbf{p}$ is the limit of a sequence of points of $\mathcal{K}$, and $\mathcal{K}$ is closed, so $\mathbf{p} \in \mathcal{K}$. Conversely, if $f(\mathbf{p}) > 1$, then from the definition of f we have $\mathbf{p} = 1^{-1}\mathbf{p} \notin \mathcal{K}$.

Finally, we must prove f satisfies $f(\mathbf{p} + \mathbf{q}) \leq f(\mathbf{p}) + f(\mathbf{q})$ for all $\mathbf{p}, \mathbf{q}$; since the proof is otherwise trivial, assume $\mathbf{p} \neq \mathbf{o}$, $\mathbf{q} \neq \mathbf{o}$, $\mathbf{p} + \mathbf{q} \neq \mathbf{o}$. Let ε be a positive number, and take

$$r = (1 + \varepsilon)f(\mathbf{p}) > 0,$$

$$s = (1 + \varepsilon)f(\mathbf{q}) > 0.$$

Then $f(r^{-1}\mathbf{p}) < 1$ and $f(s^{-1}\mathbf{q}) < 1$; therefore, $r^{-1}\mathbf{p}$ and $s^{-1}\mathbf{q} \in \mathcal{K}$, and by

convexity

$$\left(\frac{r}{r+s}\right)r^{-1}\mathbf{p} + \left(1 - \frac{r}{r+s}\right)s^{-1}\mathbf{q} = \frac{\mathbf{p} + \mathbf{q}}{r+s} \in \mathcal{K}.$$

Thus,

$$f\left(\frac{\mathbf{p} + \mathbf{q}}{r+s}\right) \leq 1,$$

or

$$f(\mathbf{p} + \mathbf{q}) \leq r + s = (1 + \varepsilon)\{f(\mathbf{p}) + f(\mathbf{q})\}.$$

But this last inequality holds for every $\varepsilon > 0$, so $f(\mathbf{p} + \mathbf{q}) \leq f(\mathbf{p}) + f(\mathbf{q})$.

EXERCISES

- 47-1. Prove the union of any number of open sets is open, the union of a finite number of closed sets is closed. Take complements and obtain corresponding results for intersections.
- 47-2. Prove the intersection of any number of convex sets is convex, and the intersection of bounded sets is bounded. What can be said about the union of convex sets? of bounded sets?
- 47-3. If $f(\mathbf{p}) = f((p_1, p_2)) = \max(|p_1|, |p_2|)$, prove f is a csdf. Describe the set $f(\mathbf{p}) \leq 1$.
- 47-4. Decide whether each of the following is a csdf, and describe the set $f(\mathbf{p}) \leq 1$.
- $f(\mathbf{p}) = f((p_1, p_2)) = |p_1 + p_2|$
 - $f((p_1, p_2)) = |p_1|$
 - $f((p_1, p_2)) = (p_1^3 + p_2^3)^{1/3}$
 - $f((p_1, p_2)) = (p_1^4 + p_2^4)^{1/4}$
 - $f((p_1, p_2)) = (p_1^2/a^2 + p_2^2/b^2)^{1/2}$, a, b constants
- 47-5. If $\mathcal{S}$ is the parallelogram with vertices $3\mathbf{a} + 2\mathbf{b}$, $-2\mathbf{a} + 4\mathbf{b}$, $-3\mathbf{a} - 2\mathbf{b}$, and $2\mathbf{a} - 4\mathbf{b}$, what is the associated csdf?
- 47-6. Suppose $\mathcal{S}$ is a bounded region and Λ is any integral lattice. Prove $\mathcal{S}$ contains at most a finite number of points of Λ .

48. The Theorems of Minkowski

Definition. If Λ is an integral lattice and $\mathcal{S}$ is a set in $\mathbb{E}_2$, then Λ is said to be $\mathcal{S}$ -admissible if no point of Λ except possibly $\mathbf{o}$ is an interior point of $\mathcal{S}$.

For example, if $\Lambda = \Lambda(\mathbf{a}, \mathbf{b})$ and $\mathcal{S}$ is the circular disc with center at $\mathbf{o}$ and radius $\frac{1}{2}$, then Λ is $\mathcal{S}$ -admissible; if $\mathcal{S}$ has center at $\mathbf{o}$ and radius 1, Λ is still $\mathcal{S}$ -admissible; if $\mathcal{S}$ has radius 1 and center at $\frac{1}{2}\mathbf{a}$, then Λ is not $\mathcal{S}$ -admissible because $\mathbf{a}$ is an interior point of $\mathcal{S}$.

If $\mathcal{S}$ is any set in $\mathbf{E}_2$ and $\mathbf{p}$ is any point in $\mathbf{E}_2$, we use the notation $\mathcal{S} \oplus \mathbf{p}$ for the set $\{\mathbf{q} + \mathbf{p} : \mathbf{q} \in \mathcal{S}\}$. If Λ is an integral lattice and t is any real number, we define $t\Lambda = \{t\mathbf{p} : \mathbf{p} \in \Lambda\}$.

Theorem 48.1. (Minkowski.) *Let f be a csdf, Λ an arbitrary integral lattice, and*

$$\mathcal{S} = \{\mathbf{p} \in \mathbf{E}_2 : f(\mathbf{p}) \leq 1\}.$$

Then Λ is $\mathcal{S}$ -admissible if and only if there are no interior points in the intersection of any two convex bodies $\mathcal{S} \oplus \mathbf{p}, \mathbf{p} \in 2\Lambda$.

Proof. Notice that the bodies $\mathcal{S} \oplus \mathbf{p}$ are convex bodies since these are just the translations of $\mathcal{S}$, and the topological properties of $\mathcal{S}$ are invariant under translation. For the proof of the theorem, let us assume there is an interior point $\mathbf{q}$ in some two bodies $\mathcal{S} \oplus \mathbf{p}_1$ and $\mathcal{S} \oplus \mathbf{p}_2$, $\mathbf{p}_i \in 2\Lambda$, $\mathbf{p}_1 \neq \mathbf{p}_2$. Since $\mathbf{q} = \mathbf{c}_i + \mathbf{p}_i$, $\mathbf{c}_i \in \mathcal{S}$, we see that $\mathbf{q} - \mathbf{p}_1$ and $\mathbf{q} - \mathbf{p}_2$ are interior points of $\mathcal{S}$. Then by Theorem 47.1,

$$f(\mathbf{p}_1 - \mathbf{q}) < 1 \quad \text{and} \quad f(\mathbf{p}_2 - \mathbf{q}) < 1.$$

Since $\mathbf{p}_i \in 2\Lambda$, the point $\mathbf{p} = \frac{1}{2}\mathbf{p}_1 - \frac{1}{2}\mathbf{p}_2 \in \Lambda$, and $\mathbf{p} \neq \mathbf{o}$. But

$$\begin{aligned} f(\mathbf{p}) &= \frac{1}{2}f(\mathbf{p}_1 - \mathbf{p}_2) = \frac{1}{2}f((\mathbf{p}_1 - \mathbf{q}) - (\mathbf{p}_2 - \mathbf{q})) \\ &\leq \frac{1}{2}f((\mathbf{p}_1 - \mathbf{q}) + f(\mathbf{p}_2 - \mathbf{q})) < 1, \end{aligned}$$

so $\mathbf{p}$ is an interior point of $\mathcal{S}$. Therefore, Λ is not $\mathcal{S}$ -admissible.

Conversely, suppose Λ is not $\mathcal{S}$ -admissible. Then there is a point $\mathbf{q} \neq \mathbf{o}$, $\mathbf{q} \in \Lambda$, and $\mathbf{q}$ is an interior point of $\mathcal{S}$. Since both $\mathcal{S}$ and Λ are symmetric with respect to $\mathbf{o}$, the point $-\mathbf{q} \in \mathcal{S} \cap \Lambda$ and $-\mathbf{q}$ is an interior point of $\mathcal{S}$. Therefore, $2\mathbf{q} \in 2\Lambda$, $-\mathbf{q} \in \mathcal{S}$, so that $\mathbf{q} = -\mathbf{q} + 2\mathbf{q}$ is an interior point of $\mathcal{S} \oplus \mathbf{q}$, and of $\mathcal{S} \oplus \mathbf{o}$.

In preparation for proving our next theorem we must be able to find the area of the set $\mathcal{S} = \{\mathbf{p} : f(\mathbf{p}) \leq 1\}$ defined by a csdf f . We find it convenient to introduce certain parallelograms which can be used in finding the area of $\mathcal{S}$. Suppose $\Lambda = \Lambda(\mathbf{p}_1, \mathbf{p}_2)$ is an integral lattice and $\mathbf{p} \in \Lambda$, say $\mathbf{p} = t_1\mathbf{p}_1 + t_2\mathbf{p}_2$. Define the parallelogram $\Pi(\mathbf{p})$ by

$$\Pi(\mathbf{p}) = \{\mathbf{q} : \mathbf{q} = u_1\mathbf{p}_1 + u_2\mathbf{p}_2, t_i \leq u_i < t_i + 1 \quad (i = 1, 2)\}.$$

Evidently the area of $\Pi(\mathbf{p})$ is independent of $\mathbf{p}$ and is completely determined by $\mathbf{p}_1$ and $\mathbf{p}_2$. Hence for all $\mathbf{p} \in \Lambda$, $\Pi(\mathbf{p}) = \Pi(\mathbf{o}) = \Delta(\Lambda)$.

Lemma 48.1. *Suppose $\Lambda(\mathbf{p}_1, \mathbf{p}_2)$ is an integral lattice. Suppose a real-valued function g is defined on $\mathbf{E}_2$ as follows: if $\mathbf{p} = u_1\mathbf{p}_1 + u_2\mathbf{p}_2$, then*

$$g(\mathbf{p}) = \max(|u_1|, |u_2|).$$

Then g is a csdf and for every positive k , the set

$$\mathcal{T} = \{\mathbf{p} \in \mathbb{E}_2 : g(\mathbf{p}) \leq k\}$$

is a parallelogram with area $4k^2\Delta(\Lambda)$.

Proof. When we have shown that $\mathcal{T}$ is a parallelogram, it will follow by Theorem 47.2 that g is a csdf. Now, $\mathbf{p} \in \mathcal{T}$ if and only if

$$\mathbf{p} = u_1\mathbf{p}_1 + u_2\mathbf{p}_2, \quad -k \leq u_i \leq k \quad (i = 1, 2).$$

Consider the four subsets of $\mathcal{T}$ which are the parallelograms $\Pi(\mathbf{o})$, $\Pi(-k\mathbf{p}_2)$, $\Pi(-k\mathbf{p}_1 - k\mathbf{p}_2)$, and $\Pi(-k\mathbf{p}_1 + k\mathbf{p}_2)$ of the lattice $\Lambda(k\mathbf{p}_1, k\mathbf{p}_2)$. These four subsets of $\mathcal{T}$ do not overlap; moreover, their union is $\mathcal{T}$ except for certain boundary points, so $\mathcal{T}$ is a parallelogram. Since the inclusion or exclusion of boundary points does not affect area, the area of $\mathcal{T}$ is 4 times the area of $\Pi(\mathbf{o})$ of $\Lambda(k\mathbf{p}_1, k\mathbf{p}_2)$. Therefore,

$$\text{area of } \mathcal{T} = 4|\det(k\mathbf{p}_1, k\mathbf{p}_2)| = 4k^2\Delta(\Lambda(\mathbf{p}_1, \mathbf{p}_2)).$$

Theorem 48.2. (Minkowski.) Let f be a csdf, let $\mathcal{S} = \{\mathbf{p} : f(\mathbf{p}) \leq 1\}$, and let $A(\mathcal{S})$ denote the area of $\mathcal{S}$. If $\Lambda(\mathbf{p}_1, \mathbf{p}_2)$ is an $\mathcal{S}$ -admissible integral lattice, then $A(\mathcal{S}) \leq 4\Delta(\Lambda)$.

Proof. Let $g(\mathbf{p}) = \max(|u_1|, |u_2|)$, where $\mathbf{p} = u_1\mathbf{p}_1 + u_2\mathbf{p}_2 \in \mathbb{E}_2$. Since $\mathcal{S}$ is bounded, g restricted to $\mathcal{S}$ is bounded, so there exists a real c such that

$$g(\mathbf{p}) \leq c, \quad \text{if } \mathbf{p} \in \mathcal{S}.$$

Suppose $n \in \mathbb{Z}^+$ and define

$$\mathcal{S}_n = \bigcup_{\substack{\mathbf{p} \in \Lambda \\ g(\mathbf{p}) \leq n}} \{\mathcal{S} \oplus 2\mathbf{p}\}.$$

It should be clear that only a finite number of points of Λ satisfy the restriction $g(\mathbf{p}) \leq n$, so $\mathcal{S}_n$ is the union of finitely many convex bodies $\mathcal{S} \oplus 2\mathbf{p}$. In fact, there are exactly $(2n+1)^2$ points $\mathbf{p} \in \Lambda$ such that $g(\mathbf{p}) \leq n$, because $g(\mathbf{p}) \leq n$ if and only if $\mathbf{p} = t_1\mathbf{p}_1 + t_2\mathbf{p}_2$ with $-n \leq t_i \leq n$; hence there are $2n+1$ choices for t_1 and $2n+1$ choices for t_2 .

Since Λ is $\mathcal{S}$ -admissible, no two of the sets $\mathcal{S} \oplus 2\mathbf{p}$ have common interior points, so the area $A(\mathcal{S}_n)$ is the sum of the areas $A(\mathcal{S} \oplus 2\mathbf{p})$. Since $A(\mathcal{S} \oplus 2\mathbf{p}) = A(\mathcal{S})$, we have

$$A(\mathcal{S}_n) = (2n+1)^2 A(\mathcal{S}).$$

Now $\mathbf{q} \in \mathcal{S} \oplus 2\mathbf{p}$ provided $\mathbf{q} = \mathbf{q}_1 + 2\mathbf{p}$, $\mathbf{q}_1 \in \mathcal{S}$ and $\mathbf{p} \in \Lambda$, so for all $\mathbf{q} \in \mathcal{S}_n$ we have

$$g(\mathbf{q}) = g(\mathbf{q}_1 + 2\mathbf{p}) \leq g(\mathbf{q}_1) + 2g(\mathbf{p}) \leq c + 2n.$$

Then $\mathcal{S}_n$ is a subset of the parallelogram $g(\mathbf{p}) \leq c + 2n$ with area $4(c + 2n)^2 \Delta(\Lambda)$, so $A(\mathcal{S}_n)$ cannot exceed this area. Hence

$$(2n + 1)^2 A(\mathcal{S}) = A(\mathcal{S}_n) \leq 4(c + 2n)^2 \Delta(\Lambda).$$

Divide through this last inequality by n^2 , then take limits as $n \rightarrow \infty$ to complete the proof.

An easy deduction from the theorem is given in Corollary 48.2a.

Corollary 48.2a. *Let f be a csdf, $\mathcal{S} = \{\mathbf{p} : f(\mathbf{p}) \leq 1\}$, and Λ any integral lattice. If $A(\mathcal{S}) > 4 \Delta(\Lambda)$, then Λ contains an interior point of $\mathcal{S}$ different from $\mathbf{o}$.*

We can also conclude Λ contains a point $\mathbf{p}$ of $\mathcal{S}$ different from $\mathbf{o}$ if $A(\mathcal{S}) = 4 \Delta(\Lambda)$, though $\mathbf{p}$ may not be an interior point of $\mathcal{S}$. The argument depends upon the fact that there is only a finite number of points $\mathbf{p} \in \Lambda$ such that $f(\mathbf{p}) \leq k$ for every constant k , which is a geometrically obvious consequence of the boundedness of the body $f(\mathbf{p}) \leq k$. If $A(\mathcal{S}) = 4 \Delta(\Lambda)$, then for every positive ε , consider the convex body $(1 + \varepsilon)\mathcal{S}$. Since $A((1 + \varepsilon)\mathcal{S}) > 4 \Delta(\Lambda)$, by Corollary 48.2a, Λ contains an interior point $\mathbf{p} \neq \mathbf{o}$ of $(1 + \varepsilon)\mathcal{S}$ for every $\varepsilon > 0$. But the points in $2\mathcal{S}$ all satisfy $f(\mathbf{p}) \leq 2$ and there are only finitely many such $\mathbf{p} \in \Lambda$; say these points are $\mathbf{q}_1, \dots, \mathbf{q}_n$. We will show that one of the $\mathbf{q}_j \in \mathcal{S}$.

Suppose instead that $\mathbf{q}_j \notin \mathcal{S}$, $j = 1, \dots, n$. Then $1 < f(\mathbf{q}_j) \leq 2$, so if we choose

$$\delta = \min_{1 \leq j \leq n} (f(\mathbf{q}_j) - 1)$$

then $\delta > 0$. There is an interior point $\mathbf{q}_i$ in $(1 + \delta/2)\mathcal{S}$, so $\mathbf{q}_i$ is of the form $t\mathbf{p}$, $\mathbf{p} \in \mathcal{S}$ and $|t| < 1 + \delta/2$. Therefore, $f(\mathbf{q}_i) = |t|f(\mathbf{p}) \leq 1 + \delta/2$; but $\delta \leq f(\mathbf{q}_i) - 1$, so $\delta + 1 \leq f(\mathbf{q}_i) \leq 1 + \delta/2$, which is a contradiction. We have proved Corollary 48.2b.

Corollary 48.2b. *Let f be a csdf, $\mathcal{S} = \{\mathbf{p} : f(\mathbf{p}) \leq 1\}$, and Λ any integral lattice. If $A(\mathcal{S}) = 4 \Delta(\Lambda)$, then Λ contains a point of $\mathcal{S}$ different from $\mathbf{o}$.*

Example. Consider this problem: Let $f(\mathbf{p}) = f((p_1, p_2)) = |p_1| + |p_2|$, and $\mathcal{S}$ be the convex body determined by $f(\mathbf{p}) \leq 1$. Find a constant c such that every integral lattice satisfying $\Delta(\Lambda) \leq c$ contains a point $\mathbf{p} \in \mathcal{S}$, $\mathbf{p} \neq \mathbf{o}$.

The body $\mathcal{S}$ is the square with vertices $\mathbf{a}, \mathbf{b}, -\mathbf{a}, -\mathbf{b}$. Since $A(\mathcal{S}) = 2$, we want c such that $2 \geq 4c$; take $c = \frac{1}{2}$. For $\Lambda((\frac{1}{2}, \frac{1}{2}), (2, 1))$, $\Delta(\Lambda) = \frac{1}{2}$ and the point $(\frac{1}{2}, \frac{1}{2}) \in \mathcal{S}$. Notice that $(\frac{1}{2}, \frac{1}{2})$ is a boundary point of $\mathcal{S}$.

For $\Lambda((1, 1), (1, \frac{1}{2}))$, again $\Delta(\Lambda) = \frac{1}{2}$, and the point $(0, \frac{1}{2}) \in \Lambda$ and is an interior point of $\mathcal{S}$.

EXERCISES

- 48-1. Suppose $f(\mathbf{p}) = |\mathbf{p}|$ and $\mathcal{S} = \{\mathbf{p} : f(\mathbf{p}) \leq 1\}$. Decide whether the following lattices are $\mathcal{S}$ -admissible:
- $\Lambda(\mathbf{a}, \mathbf{b})$;
 - $\Lambda(\mathbf{a}, (\pi/4, 1))$;
 - $\Lambda((1/\sqrt{2}, 1/\sqrt{2}), (3, 1))$.
- 48-2. With $f(\mathbf{p}) = f((p_1, p_2)) = |p_1| + |p_2|$ and $\mathcal{S}$ determined by $f(\mathbf{p}) \leq 1$, decide whether each of the following is $\mathcal{S}$ -admissible:
- $\Lambda(\mathbf{a}, \mathbf{b})$;
 - $\Lambda(\frac{1}{2}\mathbf{a}, \mathbf{b})$;
 - $\Lambda((1, 2), (3, 2))$.
- 48-3. If $f(\mathbf{p}) = f((p_1, p_2)) = \max(|p_1|, |p_2|)$, and $\mathcal{S}$ is the set determined by $f(\mathbf{p}) \leq 1$, find a point of $\mathcal{S}$ in each of the following lattices:
- $\Lambda(\frac{1}{2}\mathbf{a}, 2\mathbf{b})$;
 - $\Lambda((1, 2), (1, 3))$;
 - $\Lambda((\frac{1}{2}, 3), (\frac{1}{4}, 3))$.
- 48-4. Suppose α, β, γ are real, $\mathbf{p} = (x, y)$, and $F(\mathbf{p}) = \alpha x^2 + 2\beta xy + \gamma y^2$. If $F(\mathbf{p}) > 0$ for all $\mathbf{p} \neq \mathbf{o}$, F is called a *positive definite binary quadratic form*. The number $\delta = \alpha\gamma - \beta^2$ is the *discriminant* of the form. Prove F is positive definite if and only if $\alpha > 0$ and $\delta > 0$. [Hint: Show that there are points $\mathbf{p}, \mathbf{q}$ such that $F(\mathbf{p}) = \alpha$, $F(\mathbf{q}) = \delta$; show also that $F(\mathbf{p}) = u^2 + v^2$ for some u, v .]
- 48-5. If $F(\mathbf{p})$ is a positive definite quadratic form (see Exercise 48-4), and $f(\mathbf{p}) = \sqrt{F(\mathbf{p})}$, prove f is a csdf.
- 48-6. Suppose F , the form of Exercise 48-4, is a positive definite and f is the function defined in Exercise 48-5. Let r^2 be the least value of $F(\mathbf{p})$ for $\mathbf{p} \neq \mathbf{o}$, $\mathbf{p} \in \Lambda(\mathbf{a}, \mathbf{b})$. If $\mathcal{K} = \{\mathbf{p} : f(\mathbf{p}) \leq r\}$, $\mathbf{m} = (\sqrt{\alpha}, 0)$ and $\mathbf{n} = (\beta/\sqrt{\alpha}, \sqrt{\delta}/\sqrt{\alpha})$, prove that $\Lambda(\mathbf{m}, \mathbf{n})$ is $\mathcal{K}$ -admissible. Hence use Theorem 48.2 to conclude that the minimum value r^2 satisfies $r^2 \leq 4\sqrt{\delta}/\pi$.

49. Applications to Farey Sequences and Continued Fractions

The fractions a/b of the Farey sequence of order n can be put in a one-to-one correspondence with certain lattice points in the closed triangle $\mathcal{T}$ bounded by the lines $y = x$, $y = 0$, and $x = n$. The point (b, a) in $\mathcal{T}$ is the correspondent of the fraction a/b of $\mathcal{F}_n$ if and only if the line segment from $\mathbf{o}$ to (b, a) contains no point of $\Lambda(\mathbf{a}, \mathbf{b})$ except $\mathbf{o}$ and (b, a) . The order of occurrence of the fractions a/b in $\mathcal{F}_n$ is the same as the order in which the points (b, a) will be intersected by a line moving counter clockwise starting from the x -axis.

Suppose $\mathbf{p} = (b, a)$ and $\mathbf{q} = (d, c)$, and $a/b < c/d$ are successive in $\mathcal{F}_n$. Then there are no lattice points in the triangle $\mathbf{o p q}$ except the vertices; hence by Exercise 46-6, $\mathbf{p, q} \simeq \mathbf{a, b}$. It follows that $|\det(\mathbf{a b})| = |\det(\mathbf{p q})| = 1$, which is a restatement of Corollary 33.1a.

In the infinite wedge bounded by $\mathbf{o p}$ and $\mathbf{o q}$, let $\mathbf{m} = (v, u)$ be the interior point of minimal length with relatively prime coordinates. Clearly, $\mathbf{m}$ is the point corresponding to the median of a/b and c/d . Now neither of the triangles $\mathbf{o p m}$, $\mathbf{o m q}$ contains interior points from Λ , so $\Lambda(\mathbf{p, m}) = \Lambda(\mathbf{m, q})$ and $|\det(\mathbf{p m})| = |\det(\mathbf{m q})|$; this is another proof of Corollary 33.1b.

We now turn our attention to infinite continued fractions. If ρ is irrational and $C_n = P_n/Q_n$ are the convergents to ρ , define a sequence of vectors $\mathbf{c}_n = (Q_n, P_n)$. The geometric interpretation of $C_0 < C_2 < \dots < \rho < \dots < C_3 < C_1$ is that the points $\mathbf{c}_{2k}$ are all below the line $y = \rho x$ and the points $\mathbf{c}_{2k+1}$ are all above the line. Since $|P_n| < |P_{n+1}|$ and $|Q_n| < |Q_{n+1}|$, the vectors $\mathbf{c}_n$ increase in length as n increases, and get arbitrarily close to the line $y = \rho x$ because $\lim C_n = \rho$. In Figure 9 the situation is illustrated for $\rho = (3 + \sqrt{5})/2 = \|2, 1, 1, 1, \dots\|$.

If there were an interior point $\mathbf{p} = (b, a)$ in the triangle $\mathbf{o c}_n \mathbf{c}_{n+1}$, then a/b would be a better approximation to ρ than the convergents, and this would contradict Lemma 35.4. Hence, for every $n \in \mathcal{Z}^+$, $\Lambda(\mathbf{c}_n, \mathbf{c}_{n+1}) = \Lambda(\mathbf{a, b})$, and the area of every triangle $\mathbf{o c}_n \mathbf{c}_{n+1}$ is $1/2$. For $n \in \mathcal{Z}^+$, we have $\mathbf{c}_n, \mathbf{c}_{n+1} \simeq \mathbf{c}_n, \mathbf{c}_{n-1}$, so from Exercise 46-3, there is an integer a such that $\mathbf{c}_{n+1} \pm \mathbf{c}_{n-1} = a \mathbf{c}_n$. By considering lengths of these vectors, we can see we must take the minus sign if $a > 0$, and this leads to the recursive relations on the sequences $\{P_n\}$ and $\{Q_n\}$. The above geometric observations were made by Felix Klein.

We now prove a continued fraction is periodic if and only if it represents a quadratic irrational. First, assume ρ has a periodic expansion, say

$$\rho = \|a_0, a_1, \dots, a_n, \overline{b_1, \dots, b_m}\|,$$

where the vinculum indicates the repeating part, and we will prove ρ is a quadratic irrational. Of course, this is the trivial half of the theorem and a proof is easily obtained by following the pattern used in Exercise 32-3; if $\sigma = \|\overline{b_1, \dots, b_m}\|$ is the purely periodic part of the expansion, then

$$\sigma = \|b_1, \dots, b_m, \sigma\| = \frac{\sigma P_m + P_{m-1}}{\sigma Q_m + Q_{m-1}}$$

leads to a quadratic equation in σ with integer coefficients, so σ is a quadratic irrational; then $\rho = \|a_0, \dots, a_n, \sigma\|$ gives

$$\rho = \frac{\sigma P_n + P_{n-1}}{\sigma Q_n + Q_{n-1}}$$

and by rationalizing the denominator one completes the proof.

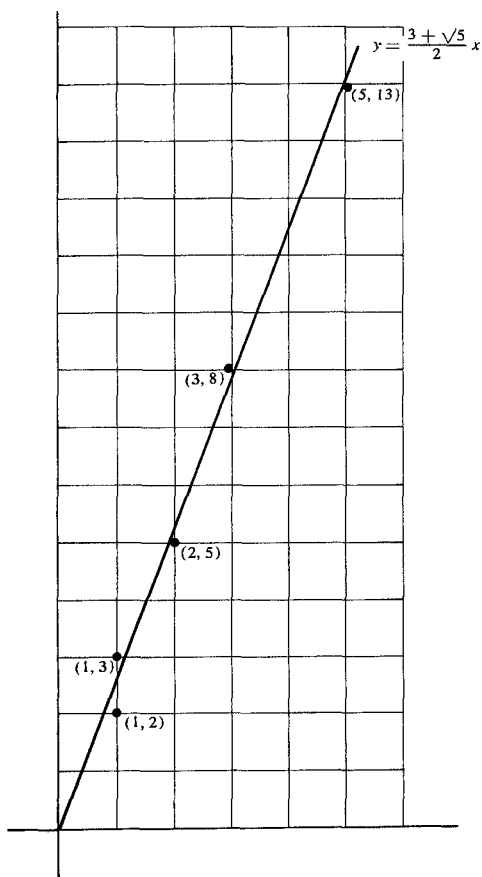


Figure 9

We prefer here to give a geometric proof that the purely periodic part σ is a quadratic irrational. In the sequence of convergents to σ , consider the subsequence $\{C'_k\}$ defined by

$$C'_1 = \|b_1, \dots, b_m\|$$

$$C'_k = \|b_1, \dots, b_m, C'_{k-1}\|, \quad k > 1.$$

The terms of this subsequence satisfy the relation

$$C'_{k+1} = \frac{P_m C'_k + P_{m-1}}{Q_m C'_k + Q_{m-1}}, \quad k = 1, 2, \dots$$

We assume $P_m Q_{m-1} - P_{m-1} Q_m = 1$; the details of the argument are not much different if this is -1 . Also, for notational ease, let $P_m = \alpha$, $P_{m-1} = \beta$, $Q_m = \gamma$, $Q_{m-1} = \delta$, and define the transformation T for all positive real x by

$$T(x) = \frac{\alpha x + \beta}{\gamma x + \delta}, \quad \alpha\delta - \beta\gamma = 1, \quad \alpha, \beta, \gamma, \delta \in \mathcal{Z}^+.$$

If we further define the iterates of T to be $T^2(x) = T(T(x))$, $T^k(x) = T(T^{k-1}(x))$, then we have

$$T^k(C'_1) = C'_{k+1}.$$

It is easily shown that if $x_1 < x_2$, then $T(x_1) < T(x_2)$. Also, $T(0) = \beta/\delta$ and $\lim_{x \rightarrow \infty} T(x) = \alpha/\gamma$; therefore $\beta/\delta \leq T(x) < \alpha/\gamma$ for all real x . In particular, regardless of the value C'_1 , the number C'_2 is on the interval $[\beta/\delta, \alpha/\gamma)$. Furthermore, there is a *fixed point* [a number x_0 such that $x_0 = T(x_0)$] under this transformation,

$$x_0 = \frac{\alpha - \delta + \sqrt{(\delta - \alpha)^2 + 4\beta\gamma}}{2\gamma} = \frac{\alpha - \delta + \sqrt{(\alpha + \delta)^2 - 4}}{2\gamma}.$$

Notice that x_0 is real because $\beta, \gamma \in \mathcal{Z}^+$, and irrational because the discriminant $(\alpha + \delta)^2 - 4$ can be a square only if $\alpha = \delta = 1$, which is not possible ($\alpha = P_m > 1$). From the equation $x_0 = T(x_0)$ it is clear that x_0 is a quadratic irrational; we will show $\sigma = x_0$ (see Figure 10).

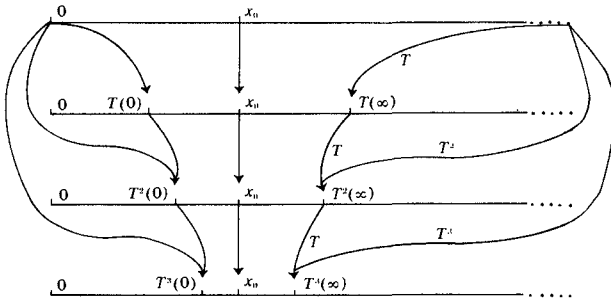


Figure 10

Consider now the difference $|T(x) - x_0|$ for any x . We have

$$\begin{aligned} |T(x) - x_0| &= |T(x) - T(x_0)| \\ &= \left| \frac{x - x_0}{(\gamma x + \delta)(\gamma x_0 + \delta)} \right| < \frac{|x - x_0|}{\gamma x_0 + \delta} < \frac{|x - x_0|}{2}. \end{aligned}$$

In the inequality $|T(x) - x_0| < |x - x_0|/2$, we replace x by $T(x)$ and continue inductively to obtain

$$|T^k(x) - x_0| < \frac{|x - x_0|}{2^k}, \quad k = 1, 2, \dots$$

With $x = C'_1$ and k sufficiently large, this proves $\{C'_k\} \rightarrow x_0$; but the sequence of all convergents approaches σ , and every subsequence must converge to the same limit. Therefore, $\sigma = x_0$ is a quadratic irrational and the conclusion is extended to ρ , as indicated previously.

Suppose conversely that ρ satisfies the equation

$$(8) \quad ax^2 + bx + c = 0, \quad a, b, c \in \mathcal{Z}$$

so that

$$\rho = \frac{-b \pm \sqrt{D}}{2a}$$

with $D = b^2 - 4ac > 0$, and D is not the square of an integer. We showed in Section 32 that if P_n/Q_n is the n^{th} convergent to ρ , then the n^{th} complete quotient θ_n^{-1} satisfies

$$\rho = \frac{P_n \theta_n^{-1} + P_{n-1}}{Q_n \theta_n^{-1} + Q_{n-1}}.$$

By solving this equation for θ_n^{-1} and using the values of ρ we find

$$\theta_n^{-1} = \frac{\rho(-Q_{n-1}) + P_{n-1}}{\rho Q_n - P_n} = \frac{\mp Q_{n-1} \sqrt{D} + 2aP_{n-1} + bQ_{n-1}}{\pm Q_n \sqrt{D} - (2aP_n + bQ_n)}.$$

We rationalize the denominator and make some algebraic simplifications to find that

$$(9) \quad \theta_n^{-1} = \pm \frac{(-1)^{n+1} \sqrt{D} - R_n}{2S_n}$$

where

$$R_n = 2aP_{n-1}P_n + b(P_{n-1}Q_n + P_nQ_{n-1}) + 2cQ_{n-1}Q_n$$

$$S_n = aP_n^2 + bP_nQ_n + cQ_n^2.$$

Now we set up a correspondence between the numbers θ_n^{-1} in (9) and points $\mathbf{p}_n \in \mathbf{E}_2$ by

$$\theta_n^{-1} \leftrightarrow \mathbf{p}_n = ((-1)^{n+1} \sqrt{D} - R_n, 2S_n).$$

Although it appears that the R_n and S_n increase without bound in absolute value as n increases, we will show that the collection of points $\mathbf{p}_n$ is a finite set.

We let $\delta_n = P_n - \rho Q_n$; then

$$|\delta_n| = Q_n \left| \frac{P_n}{Q_n} - \rho \right| < \frac{1}{Q_n}$$

by Corollary 32.1; therefore we have $P_n = \rho Q_n + \delta_n$, $|\delta_n| < 1/Q_n$. Therefore,

$$\begin{aligned} S_n &= a(\rho Q_n + \delta_n)^2 + b(\rho Q_n + \delta_n)Q_n + cQ_n^2 \\ &= (2a\rho + b)\delta_n Q_n + a\delta_n^2 \end{aligned}$$

where we have used (8) to eliminate the terms containing Q_n^2 . Taking absolute values, we find

$$|S_n| \leq |2a\rho + b| + \frac{|a|}{Q_n^2} \leq |2a\rho + b| + |a|;$$

notice that this bound is independent of n . Next we observe that $R_n^2 - 4S_{n-1}S_n = D$ for all n , and since S_n and S_{n-1} have the same upper bound we see that

$$|R_n^2| = |D + 4S_{n-1}S_n| \leq D + 4(|2a\rho + b| + |a|)^2,$$

and again the bound is independent of n .

Suppose $\mathcal{A}$ is the set of all points $\mathbf{p}_n$; let $\lambda \in \Lambda(\mathbf{a}, \mathbf{b})$ be the set $\lambda = \{-R_n \mathbf{a} + 2S_n \mathbf{b}\}$. Then

$$\mathcal{A} \subset \{\lambda \oplus (\sqrt{D}, 0)\} \cup \{\lambda \oplus (-\sqrt{D}, 0)\}.$$

Since there are only finitely many choices for the R_n and S_n , λ is a finite set and so is the set of translations $\lambda \oplus (\pm\sqrt{D}, 0)$; hence $\mathcal{A}$ contains only a finite number of points $\mathbf{p}_n$, so there are only finitely many distinct complete quotients θ_n^{-1} . It is then clear from the continued fraction algorithm that the expansion is periodic. We have proved Theorem 49.1.

Theorem 49.1. *A real number is a quadratic irrational if and only if its continued fraction expansion is infinite and periodic.*

EXERCISES

- 49-1. Supply the details for the proofs of Corollary 33.1a and Corollary 33.1b following the indications given in this section.
- 49-2. If $\mathbf{c}_n$ is the point corresponding to the n^{th} convergent C_n in the expansion of ρ , we obtained the relation $\mathbf{c}_{n+1} \pm \mathbf{c}_{n-1} = a\mathbf{c}_n$, $a \in \mathcal{L}$. Prove that the $+$ sign cannot be used, and that $a = a_n$, the n^{th} partial quotient of ρ .

49-3. Suppose $\alpha, \beta, \gamma, \delta \in \mathcal{Z}^+$, $\alpha\delta - \beta\gamma = -1$, and

$$T(x) = \frac{\alpha x + \beta}{\gamma x + \delta}, \quad x > 0.$$

Is this transformation monotonic? Is there a fixed point under T ? If x_1 is arbitrary but fixed, does the sequence $\{T^k(x_1)\}$ converge?

SOLUTIONS FOR SELECTED EXERCISES

Section 1

1-1. (a) $a \cdot 1 = a$.

(b) There exist $c, d \in \mathcal{F}$ such that $ac = b$ and $bd = a$. Therefore, $a = bd = acd$, so $cd = 1$, and $c = d = \pm 1$. From $a = bd$, we have $a = \pm b$.

(c) There exist $d, e \in \mathcal{F}$ such that $ad = b$ and $be = c$. Therefore, $c = (ad)e = a(de)$.

1-3. (a) Follows immediately from Theorem 1.1(a).

(b) Suppose $r|s$ and $s|r$. By Theorem 1.1(b), $r = \pm s$. But both r and s are positive, so $r = s$.

(c) Immediate from Theorem 1.1(c).

1-4. Suppose $a|b_1$ and x_1 is any integer. For some $d_1 \in \mathcal{F}$, $ad_1 = b_1$. Thus, $a(d_1x_1) = b_1x_1$, so $a|b_1x_1$. Assume that if $a|b_i$ ($i = 1, \dots, n$) and $\{x_1, \dots, x_n\}$ is any subset of $\mathcal{F}$, then $a|(b_1x_1 + \dots + b_nx_n)$. Now suppose $a|b_i$ ($i = 1, \dots, n+1$) and $\{x_1, \dots, x_{n+1}\} \subset \mathcal{F}$. By the inductive assumption $a|(b_1x_1 + \dots + b_nx_n)$; say $ad_1 = b_1x_1 + \dots + b_nx_n$. Also $a|b_{n+1}$; say $ad_2 = b_{n+1}$. Then $b_1x_1 + \dots + b_{n+1}x_{n+1} = a(d_1 + d_2x_{n+1})$, so a divides the required linear combination of the b_i .

1-5. (a) This is a special case of Exercise 1-4 with $n = 1$.

(b) If $ad = b$, then $(ac)d = bc$.

(c) If $ax = b$ and $cy = d$, then $(ac)xy = bd$.

1-6. Suppose $a \neq 0$. If $d|a$, then $|d| \leq |a|$, so there are at most $2|a| + 1$ choices for d .

1-7. (a) If $d \in \mathcal{D}$, then $dx = n$ for some $x \in \mathcal{F}^*$. First, $dx = n$ says $x|n$; second, $x = n/d \in \mathcal{F}^*$, so n/d is a positive divisor of n , hence is in $\mathcal{D}$. Conversely, if $n/d \in \mathcal{D}$, then $n/d > 0$ (so $d > 0$) and d is a divisor of n .

1-9. (a) Suppose m and n are even. There are integers x and y such that $m = 2x$ and $n = 2y$. Then $m \pm n = 2(x \pm y)$ is even.

(b) From Corollary 1.3 (with $b = 2$) we see that if m and n are odd, there exist q_i ($i = 1, 2$) such that $m = 2q_1 + 1$ and $n = 2q_2 + 1$. Then $m + n = 2(q_1 + q_2 + 1)$ and $m - n = 2(q_1 - q_2)$. Both are even.

(c) Suppose m is even and n is odd, and suppose $m \pm n = a$. If a is even, then $\pm n = a - m$; but n is odd, and $a - m$ is even. Thus, a cannot be even.

1-11. If $n|a$ and $n|b$, say $nx = a$ and $ny = b$. Then $0 < |a - b| = |n| |x - y| < |n|$. This implies $|x - y| < 1$, which is not possible since $x \neq y$.

1-13. If $acx = bc$ and $c \neq 0$, then $ax = b$, so $a|b$.

Section 2

2-1. (a) If $d = (a, b)$, then $d|a$, so $|c|d|a|c|$; also, $d|b$, so $|c|d|bc$. Therefore, $|c|d$ divides (ac, bc) .

Let $\mathcal{J}_2 = \{|c|n : n = ax + by, x, y \in \mathcal{J}\}$ and $\mathcal{J}_3 = \{n : n = acx + bcy, x, y \in \mathcal{J}\}$. If $k \in \mathcal{J}_2$, then $k = |c|(ax + by) = ac(\pm x) + bc(\pm y) \in \mathcal{J}_3$, so $\mathcal{J}_2 \subset \mathcal{J}_3$. We know $|c|d \in \mathcal{J}_2$ because d has a representation $d = au + bv$, so $|c|d \in \mathcal{J}_3$ and $|c|d > 0$. Since (ac, bc) is the least positive element of $\mathcal{J}_3$, $(ac, bc) \leq |c|d$. But $|c|d$ divides (ac, bc) , so $|c|d \leq (ac, bc)$.

(b) Say $ax = c$ and $by = c$. Since $ax = by$, $a|by$. But $(a, b) = 1$ implies $a|y$; say $az = y$. Thus, $c = by = (ab)z$.

(c) Since the gcd divides both entries, $|a||b|$. If $ad = b$, use Exercise 2-1(a).

(d) Say $au + bv = 1$ and $ax + cy = 1$ for some integers u, v, x, y . Then $1 = a(au + bvx + cy) + bc(vy)$, so $(a, bc) = 1$ by Corollary 2.1(b).

2-3. (a) $|a||b| = |ab| = (a, b)(a, b) = (a, c)(a, c) = |ac| = |a||c|$.

(c) Let $(b, c) = d$ and $(a, d) = e$; say $b = dB$, $c = dC$, $a = eA$, $d = eD$, and $(B, C) = 1 = (A, D)$. Then $(a, (b, c)) = (eA, (eDB, eDC)) = (eA, eD(B, C)) = e(A, DBC) = e(A, BC) = e(A, B)(A, C) = e((A, B), (A, C)) = e((A, DB), (A, DC)) = (e(A, DB), e(A, DC)) = ((eA, eDB), (eA, eDC)) = ((a, b), (a, c))$.

2-4. Since the intersection contains 1, $\mathcal{J}_1 \cap \mathcal{J}_2 \neq \emptyset$. Clearly the intersection contains all common divisors of a and b , and contains the gcd. If there is an element $e \in \mathcal{J}_1 \cap \mathcal{J}_2$ such that $(a, b) < e$, then $e|a$ and $e|b$, but $e \nmid (a, b)$. Hence (a, b) is the largest.

2-5. Say $f(x) = c_n x^n + \dots + c_1 x + c_0$. Since a/b is a zero of $f(x)$, we have $f(a/b) = 0$. Multiplying through the equation $f(a/b) = 0$ by b^n , we have $c_n a^n + c_{n-1} a^{n-1} b + \dots + c_1 a b^{n-1} + c_0 b^n = 0$. Since a obviously divides every term in the last equation except $c_0 b^n$, it follows that $a|c_0 b^n$. But $(a, b) = 1$, so $(a, b^n) = 1$, and we have $a|c_0$. Similarly, $b|c_n a^n$ and $(b, a^n) = 1$ imply that $b|c_n$.

2-7. The smallest positive element in the intersection is precisely the m which was chosen in the first paragraph after the definition of lcm.

2-8. Let $d = (a_1, a_2, a_3)$ and $e = ((a_1, a_2), a_3)$. Now $e|(a_1, a_2)$ and $e|a_3$, so $e|a_i$ ($i = 1, 2, 3$); thus, $e|d$. On the other hand, $d|a_1$ and $d|a_2$, so $d|(a_1, a_2)$; also $d|a_3$, so $d|e$. Therefore the result is true for $n = 3$.

Assume that $(a_1, \dots, a_n) = ((a_1, \dots, a_{n-1}), a_n)$ for all $n < k$ ($k \geq 4$). Let $d = (a_1, \dots, a_k)$ and $e = ((a_1, \dots, a_{k-1}), a_k)$. By the inductive hypothesis, $e = (((a_1, \dots, a_{k-2}), a_{k-1}), a_k)$. Now

$d|a_i$ ($i = 1, \dots, k$). In particular, $d|a_i$ ($i = 1, \dots, a_{k-2}$) says $d|(a_1, \dots, a_{k-2})$; also $d|a_{k-1}$, so $d|((a_1, \dots, a_{k-2}), a_{k-1})$, and $d|a_k$, so $d|e$. But $e|(a_1, \dots, a_{k-1})$ and $e|a_k$, so $e|a_i$ for $(i = 1, \dots, a_k)$ and $e|d$.

2-9. This is not significantly different from 2-8.

2-11. $(a, b)|b$, $r = a - bq$, (a, b) divides both terms on the right-hand side, so $(a, b)|r$. Therefore, $(a, b)|(b, r)$. But $(b, r)|b$ and (b, r) divides $a = bq + r$, so $(b, r)|(a, b)$.

2-13. It remains only to show that $\mathcal{S}'$ is closed under addition. Since $\mathcal{S}' \neq \phi$, there is an element $a \in \mathcal{S}'$. Hence $a - a = 0 \in \mathcal{S}'$. Suppose $b, c \in \mathcal{S}'$. Since $0 \in \mathcal{S}'$, $0 - c = -c \in \mathcal{S}'$; then $b - (-c) = b + c \in \mathcal{S}'$.

$$2-15. \quad 2 = (2, 4) = 2(3) + 4(-1) = 2(-5) + 4(3).$$

2-17. We show that $(3, n^2 + 1) = 1$ for every $n \in \mathcal{J}$. Since the only positive divisors of 3 are 1 and 3, when we have shown that $3 \nmid (n^2 + 1)$, we will be through. Let $n \in \mathcal{J}$. We use the division algorithm to find q, r such that $n = 3q + r$, $r \in \{0, 1, 2\}$. If $r = 0$, then $n^2 + 1 = (3q)^2 + 1 = 3(3q^2) + 1$; if $r = 1$, $n^2 + 1 = 3(3q^2 + 2q) + 2$; if $r = 2$, $n^2 + 1 = 3(3q^2 + 4q + 1) + 2$. Since none of those leaves a remainder 0 when divided by 3, none is a multiple of 3.

2-19. We may assume $0 < b \leq a$. Find q, r such that $a = bq + r$, $0 \leq r < b$. We will prove by mathematical induction that $(n^a - 1, n^b - 1) = n^{(a,b)} - 1$. Before making the induction, consider the algebraic identity

$$n^a - 1 = (n^b - 1)(n^{a-b} + n^{a-2b} + \dots + n^{a-\phi b})(n^b - 1).$$

Using arguments analogous to those used in Exercise 2-11, we note that $(n^a - 1, n^b - 1) = (n^b - 1, n^r - 1)$.

We use induction on a , and anchor the induction at $a = 2$. Since $0 < b \leq a$, we must verify the result for $b = 1$ and $b = 2$; those verifications are straight forward. Assume that the result holds for all a , $2 \leq a \leq A - 1$ and for all b , $0 < b \leq a$.

Suppose $a = A$. If $b = a$, the result is obvious, so assume $b < a$. With r as above, we have $0 \leq r < b \leq A - 1$, so the inductive assumption says the result is true for b and r . Thus, $n^{(a,b)} - 1 = n^{(b,r)} - 1 = (n^b - 1, n^r - 1) = (n^a - 1, n^b - 1)$. (We used Exercise 2-11 in the first equality of the chain.)

Use Theorem 2.4 for the lcm.

Section 3

$$3-1. \quad 5, 16, \text{ and } 9.$$

3-3. Use Theorem 2.4 and Exercise 3-1.

$$\begin{aligned} 3-5. \quad \langle 60, 30 \rangle &= (60)(30)/(60, 30) = 60; \\ \langle 60, 30, 42 \rangle &= \langle \langle 60, 30 \rangle, 42 \rangle = \langle 60, 42 \rangle = (60)(42)/(60, 42) = 420; \\ \langle 60, 30, 42, 8 \rangle &= \langle 420, 8 \rangle = (420)(8)/(420, 8) = 960. \end{aligned}$$

3-6. Assume $m < n$. We first prove that for any even positive integer x ,

$(x^{2^{n-m}} + 1, x + 1) = 1$. The equations in the Euclidean algorithm are

$$x^{2^{n-m}} + 1 = (x + 1) \sum_{i=1}^{2^{n-m}-1} (-1)^{i-1} x^i + 2 \quad \text{and} \quad x + 1 = 2(x/2) + 1. \quad \text{Now take}$$

$$x = (2k)^{2^m}.$$

Section 4

4-1. For every prime p , $p^\beta | p^\alpha$ if and only if $0 \leq \beta \leq \alpha$.

4-2. Let $c = \prod p_i^{\gamma_i}$. Since $\gamma_i \leq \alpha_i$, $c|a$; since $\gamma_i \leq \beta_i$, $c|b$. If $n|a$ and $n|b$, say $n = \prod p_i^{\delta_i}$; then $0 \leq \delta_i \leq \alpha_i$ and $0 \leq \delta_i \leq \beta_i$. But then $\delta_i \leq \min(\alpha_i, \beta_i) = \gamma_i$ ($i = 1, \dots, r$), so $n|c$.

The product ab has exponents $\alpha_i + \beta_i$ in its canonical form. Clearly $\alpha_i + \beta_i = \max(\alpha_i, \beta_i) + \min(\alpha_i, \beta_i)$.

4-3. Say $m = \pm \prod_{i=1}^M p_i^{\alpha_i}$ and $n = \pm \prod_{i=M+1}^N p_i^{\alpha_i}$. Since $(m, n) = 1$, $\{p_i: i = 1, \dots, M\} \cap \{p_i: i = M+1, \dots, N\} = \emptyset$. Since $d|mn$, $d = \prod_{i=1}^N p_i^{\gamma_i}$, and $0 \leq \gamma_i \leq \alpha_i$ ($i = 1, \dots, N$). Define $a = \prod_{i=1}^M p_i^{\gamma_i}$ and $b = \prod_{i=M+1}^N p_i^{\gamma_i}$. Thus, $d = ab$, $a|m$, $b|n$, and $(a, b) = 1$. Furthermore, $\frac{m}{a} = \prod_{i=1}^M p_i^{\alpha_i - \gamma_i}$ and $\frac{n}{b} = \prod_{i=M+1}^N p_i^{\alpha_i - \gamma_i}$, so $(m/a, n/b) = 1$. It is obvious that we have chosen $a = (d, m)$ and $b = (d, n)$.

For the proof of uniqueness, suppose there is another factorization $d = AB$ with the required properties. Since $ab = AB$, $a|AB$. But $a|m$, $B|n$, and $(m, n) = 1$ imply that $(a, B) = 1$, so $a|A$. Similarly, $A|ab$ and $(A, b) = 1$ imply that $A|a$, so $a = A$. Hence, $b = B$, and the choice is unique.

4-5. Since $\alpha_j \leq \delta_i$ ($j = 1, \dots, k$), $a_j | \prod p_i^{\delta_i}$ for each j . If $n = \prod p_i^{\gamma_i}$ is any integer which is a multiple of all a_j ($j = 1, \dots, k$), then $\gamma_i \geq \alpha_j$ for all i , so $\gamma_i \geq \delta_i$. Thus, $n | \prod p_i^{\delta_i}$, so $\prod p_i^{\delta_i} = \langle a_1, \dots, a_k \rangle$.

4-7. There is a prime p which divides m , because $m > 1$. If for some i , $1 \leq i \leq n$, we have $p = p_i$, then $p|m$ and $p|p_i$, which implies $p|1$.

4-9. (a) Say $a = \prod p_i^{\alpha_i}$, $b = \prod p_i^{\beta_i}$, $c = \prod p_i^{\gamma_i}$. Hint: verify that $\max(\alpha_i, \min(\beta_i, \gamma_i)) = \min(\max(\alpha_i, \beta_i), \max(\alpha_i, \gamma_i))$.

4-11. The element in row r , column c in the table is $r + c + 2rc$. If k occurs, then for some r, c , we have $k = r + c + 2rc$. Then $2k + 1 = (2r + 1)(2c + 1)$; this has two factors, each > 1 , so both factors have at least one prime divisor. Thus, $2k + 1$ is not prime.

Conversely, suppose the odd integer $2k + 1$ is not prime, so it has at least two odd factors, say $2k + 1 = ab$, $a > 1$, $b > 1$. Choose $r = (a - 1)/2$, $c = (b - 1)/2$. Then k occurs in row r and column c . (Obviously, the position for occurrence is not unique.)

Section 5

5-1. Let $f \in \mathcal{A}$. Now $f \cdot \epsilon(1) = f(1)\epsilon(1) = f'(1)$. For $n > 1$, $f \cdot \epsilon(n) = f(n)\epsilon(1) + \sum_{\substack{ab=n \\ b>1}} f(a)\epsilon(b) = f(n)$. Therefore, $f \cdot \epsilon = f$ for all $f \in \mathcal{A}$.

5-3. Except for $a = 2$ and $b = 3$, all terms $f(a)g(b) = 0$. Therefore, $f \cdot g(6) = -1$ and $f \cdot g(n) = 0$ when $n \neq 6$.

5-5. Suppose $f \cdot g = g$ for every f . In particular, then, the equality holds for that f such that $f(n) = 2$ for every n . Then $f(1)g(1) = 2g(1) = f \cdot g(1) = g(1)$ implies $g(1) = 0$. Assume $g(n) = 0$ for $1 \leq n < k$. Then, $f \cdot g(k) = f(1)g(k) = g(k)$ implies $g(k) = 0$. Thus, $g = \theta$.

Section 6

$$6-1. \quad f' = f' \cdot \epsilon = f' \cdot (f \cdot g) = (f' \cdot f) \cdot g = \epsilon \cdot g = g.$$

$$6-3. \quad f(n) = 1 \text{ for } n = 1, 6, 10; = -1 \text{ for } n = 2, 3, 5, 7; = 0 \text{ for } n = 4, 8, 9.$$

$$6-5. \quad \epsilon \cdot \epsilon = \epsilon, \text{ so } \epsilon = \epsilon' \text{ by Exercise 6-1.}$$

Section 7

7-1. $\epsilon(1) = 1 \neq 0$; if $(m, n) = 1$ and $mn > 1$, then at least one of m or n is greater than 1, so both $\epsilon(mn)$ and $\epsilon(m)\epsilon(n)$ vanish.

$$\begin{aligned} 7-2. \quad (mn)^* &= (mn)^n \{ \cos(t \log mn) + i \sin(t \log mn) \} \\ &= m^n n^n \{ \cos(t \log m + t \log n) + i \sin(t \log m + t \log n) \} \\ &= m^n n^n \{ \cos(t \log m) \cos(t \log n) - \sin(t \log m) \sin(t \log n) \\ &\quad + i \sin(t \log m) \cos(t \log n) + i \sin(t \log n) \cos(t \log m) \} \\ &= m^n \{ \cos(t \log m) + i \sin(t \log m) \} n^n \{ \cos(t \log n) + i \sin(t \log n) \} \\ &= m^* n^*. \end{aligned}$$

7-3. $f(1) = 1$, $f(n) = 2$ for $n > 2$. Not multiplicative because $f(6) = 2$, but $f(2)f(3) = 4$.

Section 8

$$8-1. \quad fh = g \text{ if and only if } f = f\epsilon = f(h \cdot h') = (fh) \cdot h' = g \cdot h'.$$

$$8-3. \quad \iota_1(1) = 1. \text{ For every prime } p, \iota_1(p) = -p, \iota_1(p^\beta) = 0 \text{ if } \beta > 1.$$

$$8-5. \quad (a), (b), (d), \text{ and } (f).$$

$$\begin{aligned} 8-7. \quad & (a) F_0, F_1, F_2 \\ & (b) F_0, F_1 \\ & (c) F_0 = \iota_1, F_1 = \iota_0 \\ & (d) F_1 \text{ is the characteristic function of } \mathcal{P}^+. \end{aligned}$$

8-9. The summation is $\lambda \cdot \eta(n)$. For prime p , and $k \geq 1$, $\lambda(p^k) = (-1)^k$ and $\eta(p^k) = 2$. Therefore,

$$\begin{aligned} \lambda \cdot \eta(p^\beta) &= \sum_{\alpha=0}^{\beta-1} \lambda(p^\alpha) \eta(p^{\beta-\alpha}) + \lambda(p^\beta) \eta(1) \\ &= \sum_{\alpha=0}^{\beta-1} (-1)^\alpha 2 + (-1)^\beta \\ &= \{ 2 \text{ if } \beta \text{ is odd, } 0 \text{ if } \beta \text{ is even} \} + \\ &\quad \{ -1 \text{ if } \beta \text{ is odd, } 1 \text{ if } \beta \text{ is even} \} \\ &= 1 \text{ for all } \beta \geq 1. \end{aligned}$$

Section 9

9-1. $\Pi(\beta_j + 1)$ is odd if and only if every factor $\beta_j + 1$ is odd if and only if every β_j is even.

$$9-3. \quad \tau \cdot \mu = (\iota_0 \cdot \iota_0) \cdot \mu = \iota_0 \cdot (\iota_0 \cdot \mu) = \iota_0.$$

9-5. First prove that $\sum_{\alpha=0}^{\beta} \alpha = \beta(\beta+1)/2$ and $\sum_{\alpha=0}^{\beta} \alpha^3 = \beta^2(\beta+1)^2/4$. Then notice that the sums are precisely those which arise when evaluating $\iota_0 \cdot \iota_0 \cdot \iota_0(p^\beta)$ and $(\iota_0 \cdot \iota_0)^3 \cdot \iota_0(p^\beta)$.

$$9-7. \quad \log f(n) = \sum_{d|n} \log g(d). \text{ Inversion yields } \sum \mu(d) \log f\left(\frac{n}{d}\right) = \log g(n).$$

The term in the summation may be written as $\log f^{\mu(d)}(n/d)$. Now exponentiation gives the desired result.

9-9. If d is a divisor of $3n+2$, then $(d,3)=1$, so either $d=3q+1$ or $d=3q+2$ for some q . Moreover, if $de=3n+2$, and $d=3q+1$, then $e=3r+2$; if $d=3q+2$, then $e=3r+1$. In either case, $d+e$ is a multiple of 3. Since all divisors of $3n+2$ may be paired in this fashion, $\sigma(3n+2)$ is a multiple of 3.

$$9-11. \text{ Suppose } k(a+1)=b+1. \text{ Then } \sigma(p^b) = \frac{p^{b+1}-1}{p-1} = \frac{p^{(\alpha+1)k}-1}{p^{\alpha+1}-1} \sigma(p^\alpha),$$

and the multiplier is the integer $1+p^{\alpha+1}+p^{2(\alpha+1)}+\dots+p^{k(\alpha+1)}$. Conversely, if $\sigma(p^b)/\sigma(p^\alpha)$ is an integer, then $(a+1)|(b+1)$.

$$9-13. \text{ If } n > 1 \text{ and } p \text{ is any prime, } \tau(p^{n-1}) = n.$$

9-15. If n is not a square, we may pair the divisor d with the divisor n/d and the product

$$\text{is } \prod_{d < \sqrt{n}} d \left(\frac{n}{d}\right) = \prod_{d < \sqrt{n}} n. \text{ This is } n^2 \text{ if and only if there are exactly two pairs of divisors.}$$

Therefore, $\tau(n)=4$. The only such n are p^3 or pq . If n is a square, say $n=r^2$, then the divisors other than r may be paired as above. The product is

$$\prod_{\substack{d|n \\ d < r}} d \left(\frac{r^2}{d}\right) r = \prod r^2 \cdot r. \text{ Such a product is of the form } r^\beta \text{ for odd } \beta, \text{ and cannot be } n^2 = r^4.$$

Section 10

10-1. $\phi = \iota_1 \cdot \iota_0'$; ι_0 and ι_1 are completely multiplicative.

$$10-3. \quad (b) \quad \phi \cdot \sigma_s = \phi \cdot (\iota_0 \cdot \iota_s) = (\phi \cdot \iota_0) \cdot \iota_s = \iota_1 \cdot \iota_s.$$

10-5. Denote the n numbers in the set as t_j , $j=1, \dots, n$, where $t_j = k + j$. Use the division algorithm on the t_j to find $t_j = nq_j + r_j$, $0 \leq r_j < n$. No two r_j are equal, so the r_j are just $0, 1, \dots, n-1$ in some order; we know that exactly $\phi(n)$ of the r_j are relatively prime to n . But $(n, t_j) = (n, nq_j + r_j) = (n, r_j)$, so exactly $\phi(n)$ of the t_j are prime to n .

10-7. Only $n = 11$.

10-9. For prime p , $\beta > 0$

$$\begin{aligned} J_k(p^k) &= \sum_{\alpha=0}^{\beta} \iota_k(p^\alpha) \mu(p^{\beta-\alpha}) \\ &= \iota_k(p^\beta) \mu(1) + \iota_k(p^{\beta-1}) \mu(p) \\ &= p^{\beta k} - p^{(\beta-1)k}. \end{aligned}$$

10-11. Neither $p = 2$ nor $p = 3$ is possible because $2p + 1$ is prime in both cases. If there is a solution x , then x contains at most one prime, because each factor $q^\beta - q^{\beta-1}$ is even and $2p$ contains exactly one factor of 2. But the only solution of $q^\beta - q^{\beta-1} = 2p$ occurs for $q = p = 3$.

10-13. The left side is $mnd \left\{ \prod_{p|m} \left(1 - \frac{1}{p}\right) \right\} \left\{ \prod_{p|d} \left(1 - \frac{1}{p}\right) \right\}$, and the right side is

$mnd \left\{ \prod_{p|m} \left(1 - \frac{1}{p}\right) \right\} \left\{ \prod_{p|d} \left(1 - \frac{1}{p}\right) \right\}$. The factors $(1 - \frac{1}{p})$ which arise from $p|d$ appear twice on each side; those which arise from $p|d$ appear exactly once on each side.

Section 11

11-1. Since $m|(a-a)$, $a \equiv a \pmod{m}$ for every a . If $a \equiv b \pmod{m}$, then $m|(a-b)$, hence $m|(b-a)$, so $b \equiv a$. If $a \equiv b$ and $b \equiv c \pmod{m}$, then $m|(a-b)$ and $m|(b-c)$. Therefore, $m|\{(a-b) + (b-c)\}$, or $m|(a-c)$, so $a \equiv c$.

11-3. If $m|(a-b)$ and $d|m$, then $d|(a-b)$.

11-5. (b) If m is prime and $a \not\equiv 0 \pmod{m}$, then $(a, m) = 1$. Then there are integers x and y such that $ax + my = 1$. That is, $ax \equiv 1 \pmod{m}$.

(c) Assume that m is composite, say $m = MN$, $M > 1$, $N > 1$. Now $M \not\equiv 0 \pmod{m}$, but M has no inverse: if $Mx \equiv 1 \pmod{m}$, then $M|m$ and $m|(Mx-1)$ says $M|(Mx-1)$.

11-7. $2 \cdot 4 \equiv 2 \cdot 1 \pmod{6}$, $4 \not\equiv 1 \pmod{6}$.

11-9. $m_1|(b-c)$ and $m_2|(b-c)$ imply that $(m_1, m_2)|(b-c)$.

Section 12

12-1. All solutions are $x = 2 + 3t$, $y = 1 - 2t$. We want $-10 < 2 + 3t < 10$, or $-4 < t < 8/3$. The desired solutions are obtained for $t = -3, -2, -1, 0, 1, 2$.

12-3. $x = 52$, unique $\pmod{105}$.

12-5. (a) 1, 1
(b) 1, 1
(c) 1, 1 or 4, -1

Section 13

13-1. If $a \equiv p \pmod{p}$, then $a^p \equiv 0 \equiv a \pmod{p}$. If $(a, p) = 1$, then $a^{p-1} \equiv 1$, so $a^p \equiv a \pmod{p}$.

13-3. If $k^n \equiv 1 \pmod{m}$, then for some M , $mM = k^n - 1$. Therefore, the equation $kx + my = 1$ has a solution $x = k^{n-1}$, $y = -M$, so $(k, m) = 1$. Find q, r such that $\phi(m) = nq + r$, $0 \leq r < n$. If $0 < r$, we have $1 \equiv k^{\phi(m)} = (k^n)^q k^r \equiv k^r \pmod{m}$. This contradicts that n is the smallest such positive power. Therefore, $r = 0$ and $n | \phi(m)$.

13-5. $ax = a(a^{\phi(m)-1}n) \equiv n \pmod{m}$.

13-7. Since no two a_j are congruent $\pmod{m}$, it follows that no two r_j are congruent. Also, $1 = (a_j, m) = (m, r_j)$. No r_j is $0 \pmod{m}$, since $(m, r_j) = 1$. Thus the $\{r_j\}$ are $\phi(m)$ distinct positive integers relatively prime to m and $< m$.

Section 14

14-1. $c_1 = \mu$.

14-3. By Corollary 14.1, $c_n(d) = \phi(d)$. Thus, $c_n \cdot \iota_1'(n) = \phi \cdot \iota_1'(n) = \mu \cdot \iota_1 \cdot \iota_1'(n) = \mu(n)$.

14-5. $d | n$ and $d | s$ if and only if $d | (n, s)$.

14-7. (a) $c_s \cdot \tau = (c_s \cdot \iota_0) \cdot \iota_0 = \delta_s \cdot \iota_0$.

Section 15

15-1. Suppose n is composite, and not the square of a prime. Then n has a factorization $n = ab$ with $1 < a < b < n$. Both factors a and b occur in $(n-1)!$. If $n = p^2$ for some prime p , both p and $2p$ are less than n , except for $n = 4$. Therefore $2p^2 | (n-1)!$, $n \neq 4$.

15-3. If $p | g(x)h(x)$, either $p | g(x)$ or $p | h(x)$.

15-5. If n is prime and $n > 4$, then n is odd and $kn = (n-1)! + 1$, by Wilson's Theorem. Since $(n-1)! + 1$ is odd, k must be odd. Then

$$\left[\frac{(n-1)!}{n} \right] = \left[\frac{kn-1}{n} \right] = \left[k - \frac{1}{n} \right] = k-1 \text{ is even.}$$

If n is composite, then $kn = (n-1)!$ by Exercise 15-1. If n is odd, then k is even, and $[(n-1)!/n] = k$. If n is composite and even and $n \geq 10$, then 2 and $n/2$ are factors of n which occur in $(n-1)!$. Moreover, $n/2 > 4$, so 4 divides $(n-1)!/n$, and again $[(n-1)!/n]$ is even. Finally $n = 6$ and $n = 8$ are easily disposed of by computation.

15-7. That the Wilson and Fermat Theorems imply the condition is obvious. Assume that for every $a \in \mathcal{F}$ and all primes p , $a + a^p(p-1)! \equiv 0 \pmod{p}$. Since this holds for all a , it holds in particular for $a = 1$; that gives Wilson's Theorem. Now using Wilson's Theorem, we have $0 \equiv a + a^p(p-1)! \equiv a + a^p(-1) \pmod{p}$, which gives Fermat's Theorem.

Section 16

16-1. 3 (is a primitive root of 7), 2 (of 9), 2 (of 13), 6 (of 17), 2 (of 25), 2 (of 27).

16-3. $\langle t_1, t_2 \rangle / \langle \nu, \langle t_1, t_2 \rangle \rangle$.

16-4. The set contains exactly $\phi(m)$ incongruent numbers, all prime to m .

16-5. By Lemma 16.5, a is a primitive root of p^β , so $A = a + kp^\beta$ is also a primitive root of p^β . By Theorem 16.5, A is a primitive root of $p^{\beta+1}$.

16-7. If a is a primitive root of both p and p^2 , then $\{t_x\} = \{\phi(p^x); x = 1, 2, \dots\}$.

Section 17

17-1. $(-540/11) = -1$, $(540/11) = 1$, $(311, 19) = 1$.

17-3. All the statements are immediate consequences of Euler's criterion. For example, since $(ab)^n = a^n b^n$, we have $(ab/p) = (a/p)(b/p)$.

17-4. $(-3/p) = (-1)^{\frac{p-1}{2}} (3/p) = (-1)^{\frac{p-1}{2}} (p/3) (-1)^{\frac{p-1}{2}} \cdot \frac{3-1}{2} = (p/3)$. Clearly $(p/3) = 1$ if $p \equiv 1 \pmod{3}$, $= -1$ if $p \equiv 2 \pmod{3}$. These conclusions are equivalent to those stated.

17-5. From the law of quadratic reciprocity, $(p/q) = (q/p)$ if and only if $\frac{p-1}{2} \cdot \frac{q-1}{2}$ is even, which is the case if and only if at least one of p or q is $1 \pmod{4}$.

17-7. By Euler's criterion, $(-1/p) = (-1)^{\frac{p-1}{2}}$. This is $+1$ if and only if $p \equiv 1 \pmod{4}$, and -1 if and only if $p \equiv 3 \equiv -1 \pmod{4}$.

Section 18

18-1. Consider $(Df(x_0), p)$. Since p is prime, the gcd is either 1 or p .

If the gcd = 1, there is a unique solution, and gcd = 1 is equivalent to $p \nmid Df(x_0)$.

If the gcd = p , then $p \mid Df(x_0)$. There will be no solution if and only if the gcd does not divide $f(x_0)/p^\beta$; there will be p solutions if and only if $p \mid \{f(x_0)/p^\beta\}$.

18-3. There are 162 solutions $\pmod{5292}$, which arise from the intermediate solutions $x \equiv 1, 3 \pmod{4}$; $x \equiv 2, 5, 8, 11, 14, 17, 20, 23, 26 \pmod{27}$; and $x \equiv 6, 13, 19, 20, 27, 31, 34, 41, 48 \pmod{49}$. The solutions are $\{41 + 42k; k = 0, 1, \dots, 125\} \cup \{215 + 294k; k = 0, 1, \dots, 17\} \cup \{227 + 294k; k = 0, 1, \dots, 17\}$.

18-5. $x^5 \equiv x$ and $x^6 \equiv x^2 \pmod{5}$; thus $x^6 + 3 \equiv x^2 + 3 \pmod{5}$. We have $x^2 + 3 \equiv 0 \pmod{5}$ if and only if $x^2 \equiv 2 \pmod{5}$, but $(2/5) = -1$. Since $f(x) \equiv 0 \pmod{5}$ has no solution, $f(x) \equiv 0 \pmod{n}$ cannot have a solution if $5|n$.

Section 19

19-1. (2) Suppose $f = o(g)$ and $h = o(g)$. Then

$$\lim_{x \rightarrow \infty} \frac{f(x) \pm h(x)}{g(x)} = \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} \pm \lim_{x \rightarrow \infty} \frac{h(x)}{g(x)} = 0. \text{ Therefore, } f \pm h = o(g).$$

(3) Suppose $f = O(g)$ and $h = O(|f|)$; then $h = O(O(g))$. Since there exist M_1 and M_2 such that $\frac{|f(x)|}{g(x)} < M_1$ and $\frac{|h(x)|}{|f(x)|} < M_2$, we have

$$\frac{|h(x)|}{g(x)} = \frac{|h(x)|}{|f(x)|} \cdot \frac{|f(x)|}{g(x)} < M_1 M_2, \text{ so } h = O(g).$$

(5) Suppose $f = O(g)$ and $h = O(g)$; then $fh = O(g)O(g) = \{O(g)\}^2$. Since $\frac{|f(x)|}{g(x)} < M_1$ and $\frac{|h(x)|}{g(x)} < M_2$, we have $\frac{|fh(x)|}{g^2(x)} = \frac{|f(x)h(x)|}{g(x)g(x)} < M_1 M_2$, so $fh = O(g^2)$.

19-3. Suppose f, g , and h are positive-valued functions. Since $\lim_{x \rightarrow \infty} \frac{f(x)}{f(x)} = 1$, we have $f \sim f$ for all f . If $f \sim g$, then $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1$, so

$$\lim_{x \rightarrow \infty} \frac{g(x)}{f(x)} = \lim_{x \rightarrow \infty} \frac{1}{f(x)/g(x)} = 1, \text{ and } g \sim f. \text{ If } f \sim g \text{ and } g \sim h, \text{ we have}$$

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1 = \lim_{x \rightarrow \infty} \frac{g(x)}{h(x)}. \text{ Then,}$$

$$\lim_{x \rightarrow \infty} \frac{f(x)}{h(x)} = \lim_{x \rightarrow \infty} \frac{f(x)g(x)}{g(x)h(x)} = \left\{ \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} \right\} \left\{ \lim_{x \rightarrow \infty} \frac{g(x)}{h(x)} \right\} = 1, \text{ so } f \sim h.$$

19-5. $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 0 = \lim_{x \rightarrow \infty} \frac{g(x)}{h(x)}$ imply $\lim_{x \rightarrow \infty} \frac{f(x)}{h(x)} = 0$, so if $f = o(g)$ and $g = o(h)$, then $f = o(h)$. If $f(x) = 1$ for all x , then $\lim_{x \rightarrow \infty} \frac{f(x)}{f(x)} = 1 \neq 0$, so $f \neq o(f)$. If $f(x) = 1$ and $g(x) = x$, then $f = o(g)$, but $g \neq o(f)$.

19-6. We start with the formula given in Lemma 19.1. Clearly, $\log x = O(\log x)$,

$\gamma = O(\log x)$ and $O(1/x) = O(\log x)$. By property (1), we have $\sum \frac{1}{n} = O(\log x)$.

19-7. [(a) $\Rightarrow$ (b)] $O(f) = O(g)$ means $\mathcal{A}(f) \subset \mathcal{A}(g)$. That is, if $h = O(f)$, then $h = O(g)$.

[(b) $\Rightarrow$ (c)] Since $f = O(f)$, (b) says $f = O(g)$.

[(c) $\Rightarrow$ (a)] Suppose $f = O(g)$, and suppose $|f(x)|/g(x) < M_1$. Let $h \in \mathcal{O}(f)$. Then $|h(x)|/f(x) < M_2$, so $|h(x)|/g(x) < M_1 M_2$. Thus, $h = O(g)$, so $h \in \mathcal{O}(g)$. Therefore, $\mathcal{O}(f) \subset \mathcal{O}(g)$, so $O(f) = O(g)$.

19-9. Apply property (2) twice: $O(O(O(g))) = O(O(g)) = O(g)$.

Section 20

$$\begin{aligned} 20-1. \text{ Notice that } \int_1^{\infty} \frac{t - [t]}{t^2} dt &= \lim_{N \rightarrow \infty} \left\{ \int_1^N \frac{1}{t} dt - \int_1^N \frac{[t]}{t^2} dt \right\} \\ &= \lim_{N \rightarrow \infty} \left\{ \log N - \int_1^N \frac{[t]}{t^2} dt \right\}. \end{aligned}$$

$$\text{Therefore, } C_1 - \gamma = 1 - \lim_{N \rightarrow \infty} \left\{ \sum_{n=1}^N \frac{1}{n} - \int_1^N \frac{[t]}{t^2} dt \right\}.$$

$$\begin{aligned} \text{Now } \sum_{n=1}^N \frac{1}{n} - \int_1^N \frac{[t]}{t^2} dt &= 1 + \sum_{n=2}^N \left\{ \frac{1}{n} - \int_{n-1}^n \frac{[t]}{t^2} dt \right\} \\ &= 1 + \sum_{n=2}^N \left\{ \frac{1}{n} - (n-1) \int_{n-1}^n \frac{dt}{t^2} \right\} \\ &= 1 + \sum_{n=2}^N 0. \text{ Thus, } C_1 - \gamma = 0. \end{aligned}$$

20-3. The Euler - McLaurer sum formula yields

$$\sum_{n \leq x} \frac{1}{n^2} = 1 + \int_1^x \frac{dt}{t^2} - \int_1^{\infty} \frac{t - [t]}{t^3} dt + \int_x^{\infty} \frac{t - [t]}{t^3} dt + O(1/x^2). \text{ The first}$$

integral is $1 - 1/x$; the third integral is $O(1/x^2)$ (so the second integral converges); the

second integral is positive and $< \int_1^{\infty} t^{-3} dt = 1$. Thus,

$$1 < K = 2 - \int_1^{\infty} \frac{t - [t]}{t^3} dt < 2.$$

$$20-5. \quad 0 < I = \int_1^x (t - [t]) t^{k-1} dt \leq \int_1^x t^{k-1} dt = \frac{x^k - 1}{k} = O(x^k).$$

$$\text{Now } \sum_{n \leq x} n^k = 1 + \int_1^x t^k dt + I + O(x^k)$$

$$= 1 + \frac{x^{k+1}}{k+1} - \frac{1}{k+1} + O(x^k). \text{ But } 1 - \frac{1}{k+1} = O(x^k) \text{ and}$$

$$C(k) = \frac{1}{k+1}.$$

Section 22

22-1. $\lim_{x \rightarrow \infty} \frac{x \log x}{x^2} = 0$, so $x \log x = o(x^2)$. If $f = o(x^2)$, then $\lim_{x \rightarrow \infty} \frac{f(x)}{x^2} = 0$.

Thus, for every $\epsilon > 0$, there exists x_0 such that if $x > x_0$, then $|f(x)/x^2| < \epsilon$. With $\epsilon = 1$, that says that for $x > x_0$, $|f(x)/x^2| < 1$, so $f = O(x)$.

Section 23

23-1. If $|f(n)| < M$, then $\sum |f(n)|n^{-2} < M \sum n^{-2}$, and the hyperharmonic series converges for $s > 1$. Thus $\sum f(n)n^{-s}$ converges absolutely.

23-3. Pattern the proof after that of Corollary 23.2b, and find

$$\sum_{n \leq x} \frac{\mu(n)}{n^4} = \frac{1}{\zeta(4)} + O(x^{-3}).$$

$$23-5. \lim_{x \rightarrow \infty} \sum_{n \leq x} n^{-2} = \sum_{n=1}^{\infty} n^{-2} = \zeta(2); \text{ and}$$

$$\lim_{x \rightarrow \infty} \left(K - \frac{1}{x} + O\left(\frac{1}{x^2}\right) \right) = K - \lim_{x \rightarrow \infty} \frac{1}{x} + \lim_{x \rightarrow \infty} O\left(\frac{1}{x^2}\right) = K.$$

Section 25

$$25-1. 105 = 10^2 + 2^2 + 1^2 + 0^2, 385 = 19^2 + 4^2 + 2^2 + 2^2, \\ 625 = 25^2 + 0^2 + 0^2 + 0^2.$$

25-3. If x is any integer, then $x^2 \equiv 0, 1, 4 \pmod{8}$. The sum of three such numbers cannot be $\equiv 7 \pmod{8}$.

Section 26

26-1. Suppose $2^{2\alpha} = x^2 + y^2$. If $\alpha = 0$, it is obvious that $1 = (\pm 1)^2 + 0^2$ is the only representation, so assume $\alpha > 0$. If x is odd, then y is also odd. But the squares of odd numbers are $\equiv 1 \pmod{4}$, so $x^2 + y^2 \equiv 2 \pmod{4}$ and $2^{2\alpha} \equiv 0 \pmod{4}$. Thus, both x and y are even.

Say $x = 2^\beta X$ and $y = 2^\gamma Y$, with $0 < \beta \leq \gamma$ and X and Y odd. Dividing both sides by $2^{2\beta}$, we get

$$(*) \quad 2^{2(\alpha - \beta)} = X^2 + 2^{2(\gamma - \beta)} Y^2.$$

If $\alpha > \beta$, the left side of (*) is $\equiv 0 \pmod{4}$, while the right side is either $\equiv 1$ or $2 \pmod{4}$, according as $\beta < \gamma$ or $\beta = \gamma$. Therefore, $\alpha = \beta$, so the right side of (*) is 1. The only solutions are $X = \pm 1, Y = 0$ or $X = 0, Y = \pm 1$, and $\beta = \gamma = \alpha$. Both possibilities lead to the same conclusion: the only representation is $2^{2\alpha} = (\pm 2^\alpha)^2 + 0^2$.

To prove that $2^{2\alpha+1} = (\pm 2^\alpha)^2 + (\pm 2^\alpha)^2$ is the only representation of $2^{2\alpha+1}$ as a sum of two squares, we use mathematical induction. It is obvious that $2 = (\pm 1)^2 + (\pm 1)^2$ and $2^3 = (\pm 2)^2 + (\pm 2)^2$, so the result is true for $\alpha = 0$ and $\alpha = 1$.

Assume the result is true up to exponent $2(\alpha - 1) + 1$, $\alpha \geq 1$. Suppose $2^{2\alpha+1} = x^2 + y^2$. As before both x and y must be even, so let us say $x = 2^\beta X$ and $y = 2^\gamma Y$, with β and γ positive and X and Y odd. Dividing by 4, we obtain $2^{2(\alpha-1)+1} = (2^{\beta-1}X)^2 + (2^{\gamma-1}Y)^2$. The inductive assumption tells us we must have $X = \pm 1$ and $Y = \pm 1$ and $\beta = \gamma = \alpha$.

26-3. If $3^{2\beta} = x^2 + y^2$, it is easy to see that both x and y must be multiples of 3: if neither x nor y is a multiple of 3, then $x^2 + y^2 \equiv 2 \pmod{3}$, while $3^{2\beta} \equiv 0 \pmod{3}$; if only one of x , y is a multiple of 3, then $x^2 + y^2 \equiv 1 \pmod{3}$.

Now use mathematical induction and proceed as in Exercise 26-1.

Section 28

28-1. To find the gcd, we use the (Gaussian) Euclidean algorithm. With $\alpha = 3 + 5i$ and $\beta = 1 - 3i$, we find $\alpha/\beta = -6/5 + (7/5)i$. Now we choose m and n satisfying $|-6/5 - m| \leq 1/2$ and $|7/5 - n| \leq 1/2$; we take $m = -1$, $n = 1$, $\lambda = -1 + i$, and $\theta = 1 + i$. Continuing in a similar way, we find

$$3 + 5i = (1 - 3i)(-1 + i) + (1 + i)$$

$$1 - 3i = (1 + i)(-1 - 2i) + 0.$$

Therefore, $(1 - 3i, 3 + 5i) = 1 + i$.

For the prime decompositions, start by factoring out the gcd from each number. We find $(1 - 3i) = (1 + i)(-1 - 2i)$ and $3 + 5i = (1 + i)(4 + i)$. The factors are Gaussian primes since their norms are rational primes.

28-3. The result is not true. For example, suppose $\beta = -\alpha \neq 0$.

28-5. $13 + i = (2 - 4i)(1 + 2i) + (3 + i)$, $D(3 + i) = 10 < 20 = D(2 - 4i)$,
 $13 + i = (2 - 4i)(2 + 3i) + (-3 + 3i)$, $D(-3 + 3i) = 18 < 20$.

28-7. (a) Since $(\alpha, \beta) | \alpha$, there is a Gaussian integer γ such that $(\alpha, \beta)\gamma = \alpha$. Therefore $\mu = \gamma\beta$ and since $\mathcal{G}$ is closed under multiplication, $\mu \in \mathcal{G}$.

(b) Say $\mu_1 = \frac{\alpha\beta}{(\alpha, \beta)_1}$ and $\mu_2 = \frac{\alpha\beta}{(\alpha, \beta)_2}$. Since any two gcd's are

associates, for some n , $0 \leq n \leq 3$, $(\alpha, \beta)_2 = i^n(\alpha, \beta)_1$. Therefore,

$$i^n \mu_2 = \frac{\alpha\beta}{(\alpha, \beta)_1} = \mu_1.$$

(c) With $(\alpha, \beta)\gamma = \alpha$, we have $\mu = \gamma\beta$, so $\beta | \mu$. With $(\alpha, \beta)\delta = \beta$, we have $\mu = \alpha\delta$, and $\alpha | \mu$.

(d) Suppose $\alpha = \prod \pi_i^{a_i}$ and $\beta = \prod \pi_i^{b_i}$ are the factorizations of α and β into primes π_i . Then $\mu = \prod \pi_i^{a_i + b_i - \min(a_i, b_i)} = \prod \pi_i^{\max(a_i, b_i)}$. Since $\alpha | \gamma$ and $\beta | \gamma$, $\pi_i^{\max(a_i, b_i)}$ is a divisor of γ for each i , so $\mu | \gamma$.

Section 30

30-1. Notice that if $\alpha \geq 0$, $\tau_1(2^\alpha) = 1$ and $\tau_3(2^\alpha) = 0$. Now use (10) and (11) with $a = 2^\alpha$ and $b = n_1 n_2$.

30-3. By the observations in Exercise 30-1, $F(2^\alpha) = 1$. For the balance of this exercise, we use the fact that F is multiplicative. If p is prime and $p \equiv 1 \pmod{4}$, then every divisor of p^α is of the form $4m + 1$; that is $\tau_3(p^\alpha) = 0$, and it follows that $F(n_1) = \tau_1(n_1)$.

If $q \equiv 3 \pmod{4}$ and q is prime, then $\tau_1(q^{2\beta}) = \beta + 1$ and $\tau_3(q^{2\beta}) = \beta$; but $\tau_1(q^{2\beta+1}) = \beta + 1$ and $\tau_3(q^{2\beta+1}) = \beta + 1$. Thus, $F(n_2) = 1$ or 0 according as n_2 is or is not a square.

Section 31

$$\begin{aligned} 31-1. \quad C_{k+1} &= \frac{(a_k + a_{k+1}^{-1})P_{k-1} + P_{k-2}}{(a_k + a_{k+1}^{-1})Q_{k-1} + Q_{k-2}} \\ &= \frac{(a_k P_{k-1} + P_{k-2}) + a_{k+1}^{-1} P_{k-1}}{(a_k Q_{k-1} + Q_{k-2}) + a_{k+1}^{-1} Q_{k-1}} \\ &= \frac{P_k + a_{k+1}^{-1} P_{k-1}}{Q_k + a_{k+1}^{-1} Q_{k-1}} = \frac{a_{k+1} P_k + P_{k-1}}{a_{k+1} Q_k + Q_{k-1}} = \frac{P_{k+1}}{Q_{k+1}}. \end{aligned}$$

31-3. From $d_j = c_j a_{j+1} + \frac{c_j c_{j+1}}{d_{j+1}}$, using $d_j = c_{j-1}$ and $d_{j+1} = c_j$, we have $c_{j-1} = a_{j+1} c_j + c_{j+1}$. Thus, c_{j+1} is a residue of $c_{j-1} \pmod{c_j}$. But all c_j are non-negative, so $0 \leq c_{j+1}$; moreover, $c_{j+1} < d_{j+1} = c_j$, so c_{j+1} is the least non-negative residue of c_{j-1} .

Section 32

32-1. The proof is similar to that for finite continued fractions, except that here we need not worry about the last partial quotient.

32-3. $\xi = 1 + 1/\xi$ is equivalent to $\xi^2 - \xi - 1 = 0$. The quadratic formula tells us that $\xi = (1 \pm \sqrt{5})/2$, but $\xi > 0$ so the root with the minus sign is extraneous.

32-5. Since $1 + \theta_0 = \xi = 1 + 1/\xi$, $\theta_0^{-1} = \xi$. Assume $\theta_n^{-1} = \xi$. Then

$$1 + \theta_{n+1} = \theta_n^{-1} = \xi = 1 + \frac{1}{\xi}, \text{ so } \theta_{n+1}^{-1} = \xi.$$

32-6. In Exercise 32-4, it was shown that $P_n = Q_{n+1}$. Thus, the sequence

$\{Q_n / Q_{n+1}\} = \{Q_n / P_n\} = \{1 / C_n\}$. In Exercise 32-3 it was shown that the sequence of convergents has limit ξ . Thus, $\lim_{n \rightarrow \infty} \frac{Q_n}{Q_{n+1}} = \lim_{n \rightarrow \infty} \frac{1}{C_n} = \frac{1}{\xi}$.

$$\text{Since } \theta_{n+1}^{-1} = \xi, \lim_{n \rightarrow \infty} \left(\theta_{n+1}^{-1} + \frac{Q_n}{Q_{n+1}} \right) = \lim_{n \rightarrow \infty} \xi + \lim_{n \rightarrow \infty} \frac{Q_n}{Q_{n+1}} = \xi + \frac{1}{\xi} = \sqrt{5}$$

Section 33

$$33-1. \mathcal{F}_5 : \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}.$$

$$\mathcal{F}_6 : \frac{0}{1}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{1}{1}.$$

33-3. $0 \leq a \leq b \leq m$ and $(a, b) = 1$ if and only if $0 \leq b - a \leq b \leq m$ and $(b - a, b) = 1$. Moreover, $b - a = a$ if and only if $a = 1, b = 2$. Therefore, except for $1/2$, the elements in $\mathcal{E}_m$ may be added in pairs and $a/b + (b - a)/b = 1$. We get the same result for $\sum_{x \in \mathcal{E}_m} x$ if we add $1/2$ for each element in $\mathcal{E}_m$ different from $1/2$, and then

add $1/2$ (since the element $1/2$ has not been included in the sum). Thus,

$$\sum_{x \in \mathcal{E}_m} x = \frac{1}{2} + \frac{1}{2}(\nu \mathcal{E}_m - 1) = \frac{1}{2} \nu \mathcal{E}_m.$$

33-4. Since there are plenty of rationals between any two reals, choose a rational number e such that $x' < e < x''$. Say $e = a/b$, with $(a, b) = 1$. Then $e \in \mathcal{E}_b$.

Section 34

34-1. We have $x^2 - Dy^2 = 1 = u^2 - Dv^2$, or $D(v^2 - y^2) = u^2 - x^2$. Since $0 < x < u$, the right side is > 0 , so $D(v^2 - y^2) > 0$. Because $D > 0$, we know $v^2 - y^2 > 0$; since $0 < y$ and $0 < v$, this says $y < v$.

34-3. Let $\alpha = x + \sqrt{D}y$ and $\beta = u + \sqrt{D}v$. Then

$$\begin{aligned} (\alpha^{-1}\beta^{-1})^{-1} &= \frac{1}{(x - \sqrt{D}y)(u - \sqrt{D}v)} \\ &= \frac{(x + \sqrt{D}y)(u + \sqrt{D}v)}{(x^2 - Dy^2)(u^2 - Dv^2)} \\ &= \alpha\beta \in \mathcal{P}. \end{aligned}$$

34-5. Let γ be the dominant fundamental factor. By Theorem 34.2, $\mathcal{P} = \{\gamma^n: n \in \mathcal{Z}^+\}$. Consequently, $\alpha \in \{\beta^{-1}: \beta \in \mathcal{P}\}$ if and only if $\alpha = (\gamma^n)^{-1} = \gamma^{-n}$ for some $n \in \mathcal{Z}^+$. Also, $\gamma^0 = 1$, so $\mathcal{P} = \{\gamma^n: n \in \mathcal{Z}\}$.

Associativity in $\mathcal{P}$ is inherited from the associative law for multiplication of reals. The element 1 is obviously the identity. If $\alpha, \beta \in \mathcal{P}$, then $\alpha = \gamma^m$ and $\beta = \gamma^n$ for some $m, n \in \mathcal{Z}$. Then $\alpha\beta = \gamma^{m+n} \in \mathcal{P}$. Finally, if $\alpha \in \mathcal{P}$, $\alpha = \gamma^n$ for some n , and $\alpha^{-1} = \gamma^{-n} \in \mathcal{P}$.

34-7. By definition, -1 is a quadratic non-residue (mod D) means the congruence $x^2 \equiv -1 \pmod{D}$ has no solution. The existence of a solution of $x^2 - Dy^2 = -1$ would contradict that -1 is a non-residue.

Section 35

35-1. We try $1/2, 1$, and $2/3$ (the mediant). We find that $\left| \frac{\sqrt{5}-1}{2} - \frac{2}{3} \right| < \frac{1}{9\sqrt{5}}$.

Now, since $\frac{\sqrt{5}+5}{2} = \frac{\sqrt{5}-1}{2} + 3$, we write

$$\left| \frac{\sqrt{5} + 5}{2} - \frac{11}{3} \right| = \left| \frac{\sqrt{5} - 1}{2} + 3 - \left(\frac{2}{3} + 3 \right) \right| < \frac{1}{9\sqrt{5}}.$$

35-3. $1/\sqrt{5} = \llbracket 0, 2, 4, 4, 4, \dots \rrbracket$. The convergents are $0, 1/2, 4/9, 17/38, 72/161, \dots$.

Section 36

36-1. $3, 4, 5; 5, 12, 13; 7, 24, 25; 8, 15, 17; 9, 40, 41; 12, 35, 37; 20, 21, 29$.

36-3. Suppose the triangle has base x , height y , and hypotenuse z . Then x, y, z is a Pythagorean triple. If the area is 60, then $xy = 120$. Since the only factorizations of 120 into two factors x and y , both of which are members of a Pythagorean triple, is $\{x, y\} = \{8, 15\}$, the only such triangle has sides $8, 15, 17$.

36-5. Suppose $x^2 + y^2 = z^2$. Assume that none of x, y, z is a multiple of 3. Then $x^2 + y^2 \equiv 2 \pmod{3}$, but $z^2 \equiv 1 \pmod{3}$. Therefore, at least one of x^2, y^2, z^2 is divisible by 3, and since 3 is prime at least one of x, y, z is a multiple of 3.

If none of the triple is divisible by 5, then the squares are $\equiv 1, 4 \pmod{5}$. Then $x^2 + y^2 \equiv 0, 2, 3 \pmod{5}$, but $z^2 \equiv 1, 4 \pmod{5}$. Therefore, 5 divides one of the squares, so as above 5 divides one of x, y, z .

If the triple is primitive, then by Theorem 36.1, $4|x$. If the triple is imprimitive, then the triple consists of three common multiples of some primitive triple; the factor 4 would still be present.

Finally, since $1 = (3, 4) = (3, 5) = (4, 5)$, the product $3 \cdot 4 \cdot 5 = 60$ divides xyz .

Section 37

37-1. If there is a solution of $x^{4m} + y^{4m} = z^{4m}$, then $(x^m)^4 + (y^m)^4 = (z^m)^4$. That contradicts Theorem 37.1.

Section 38

38-1. Say $\alpha = a + b\xi$ and $\beta = c + d\xi$. Then $\alpha \pm \beta = (a \pm c) + (b \pm d)\xi$ and $\alpha\beta = (ac - bd) + (ad + bc + bd)\xi$.

38-3. (a) If $\alpha = a + b\xi$, then $D(\bar{\alpha}) = D(a + b\xi^2) = D((a - b) - b\xi) = (a - b)^2 - (a - b)(-b) + (-b)^2 = a^2 - ab + b^2$.

(b) $D(n\alpha) = D(na + nb\xi) = (na)^2 - (na)(nb) + (nb)^2 = n^2(a^2 - ab + b^2)$.

38-5. $408 + 306\xi = 2 \cdot 3 \cdot 17 \cdot (4 + 3\xi)$; both 2 and $17 \equiv 5 \pmod{6}$ are primes in $\mathcal{J}$. The number $3 = (1 - \xi)(2 + \xi)$, and $1 - \xi$ and $2 + \xi$ are associated primes. Finally, $4 + 3\xi = \xi(-1 - 4\xi)$, so $4 + 3\xi$ is an associate of $-1 - 4\xi$, which is prime because this is $(1 - 2) + 2(-2)\xi$ and $1^2 + 3(-2)^2 = 13 \equiv 1 \pmod{6}$. Therefore, the prime decomposition is $408 + 306\xi = \xi(2)(17)(1 - \xi)(2 + \xi)(-1 - 4\xi)$.

Next, $6 + 11\xi = (3 + 2\xi)(4 + 3\xi)$. We saw above that $4 + 3\xi$ is prime. Also $3 + 2\xi$ is prime, because $3 + 2\xi = (2 + 1) + 2(1)\xi$ and $2^2 + 3(1)^2 = 7$, a rational prime $\equiv 1 \pmod{6}$.

38-7. Suppose α is prime and $\alpha|\beta\gamma$. If $\alpha \nmid \beta$, then $(\alpha, \beta) = 1$. Then there exists ϵ, η such that $\alpha\epsilon + \beta\eta = 1$. (The Euclidean algorithm can be used to solve the equation.) We then have $\alpha\gamma\epsilon + \beta\gamma\eta = \gamma$. Since α divides the terms on the left, $\alpha|\gamma$.

Section 41

$$\begin{aligned}
 41-1. \quad \sum_{n \leq x} \log n &= \int_1^x \log t \, dt + \int_1^x \frac{t - [t]}{t} dt + O(\log x) \\
 &= (t \log t - t) \Big|_1^x + O\left(\int_1^x \frac{dt}{t}\right) + O(\log x) \\
 &= x \log x - x + 1 + O(\log x) \\
 &= x \log x - x + O(\log x).
 \end{aligned}$$

$$\begin{aligned}
 41-3. \quad \sum_{n \leq x} \log \frac{x}{n} &= \sum_{n \leq x} [\log x - \log n] \\
 &= [x] \log x - \sum_{n \leq x} \log n \\
 &= [x + O(1)] \log x - [x \log x - x + O(\log x)] \\
 &= x + O(\log x).
 \end{aligned}$$

Section 42

42-1. From Theorem 42.3 we have

$$\begin{aligned}
 (\log x) \sum_{n \leq x} \Lambda(n) &= 2 \sum_{n \leq x} \Lambda(nL)(n) + O(x \log x), \text{ and from Theorem 42.4 we have} \\
 2 \sum_{n \leq x} \Lambda(nL)(n) &= 2x \log^2 x - 2 \sum_{n \leq x} \Lambda(nL)(n) + O(x \log x). \text{ Adding the} \\
 &\text{formulas, we get the desired result.}
 \end{aligned}$$

$$\begin{aligned}
 42-2. \quad \psi(x) \log x &= \sum_{n \leq x} \Lambda(n) + O(x), \text{ by Lemma 42.1,} \\
 &= [2x \log x - \sum_{n \leq x} \Lambda(n) + O(x)] + O(x),
 \end{aligned}$$

by Theorem 41.1. Adding $\sum_{n \leq x} \Lambda(n)$ to both sides yields Selberg's formula. Then

multiplying through by $\log x$ completes the exercise.

42-3 and 42-4. Make substitutions from the indicated results.

42-5. Both sides are $= 0$ if n contains at least 3 distinct prime divisors; $= (\log p \log q)/n$ if n is $p^\alpha q^\beta$ for distinct primes p and q , and α and β are positive; and $= ((\alpha - 1) \log^2 p)/n$, if $n = p^\alpha$ for prime p , $\alpha > 0$.

Section 44

$$44-1. \quad \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1 \text{ if and only if } \lim_{x \rightarrow \infty} \left(\frac{f(x)}{g(x)} - 1 \right) = 0 \text{ if and only if}$$

$$\lim_{x \rightarrow \infty} \frac{f(x) - g(x)}{g(x)} = 0. \text{ This is equivalent to } f(x) - g(x) = o(g(x)), \text{ or } f = g + o(g).$$

Since $f \sim g$ if and only if $g \sim f$, we have $g = f + o(f)$, or $f = g + o(f)$.

Section 45

45-1. If p is prime and $\sqrt{x} < p$, then no higher power of p is $\leq x$. Thus, for $\sqrt{x} < p$, each prime makes the same contribution to $\psi(x)$ and to $\theta(x)$. Thus,

$$\theta(x) = \sum_{p \leq x} \log p = \sum_{p \leq \sqrt{x}} \log p + \sum_{\sqrt{x} < p \leq x} \log p, \text{ or } \theta(x) = \theta(\sqrt{x}) + \psi(x) - \psi(\sqrt{x}).$$

The above equality can be used to prove the equivalence of (23) and (24). First assume $\psi(x) \sim x$. This of course implies that $\psi(x) = O(x)$ and $\psi(x) = x + o(x)$. Now $\theta(\sqrt{x}) \leq \psi(\sqrt{x}) = O(\sqrt{x})$, so from the above we have $\theta(x) = O(\sqrt{x}) + x + o(x) + O(\sqrt{x})$. But $O(\sqrt{x}) = o(x)$. Therefore, $\theta(x) = x + o(x)$, which is (24).

The reverse implication can be obtained from the same equation used here, together with (8).

45-3. The equivalence of the two theorems is an immediate consequence of the transitivity of " $\sim$ ". We have $\pi(x) \sim Li(x)$ and $Li(x) \sim x/\log x$, so $\pi(x) \sim x/\log x$; thus, Theorem 45.5 implies the Prime Number Theorem. The reverse implication is just as easy.

Section 46

46-1. The necessary and sufficient condition is that $\det U = \pm 1$.

Assume $\det U = \pm 1$. Since $\det U \neq 0$, we know U^{-1} exists. In general we know that if $U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $U^{-1} = \frac{1}{\det U} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. With $\det U = \pm 1$, U^{-1} is integral.

Conversely, suppose U^{-1} exists and is integral. Then both $\det U$ and $\det U^{-1}$ are integers. Since $1 = \det I = \det(UU^{-1}) = (\det U)(\det U^{-1})$, $(\det U) | 1$, so $\det U = \pm 1$.

46-3. By Theorem 46.1, there is a matrix U such that

$$(*) \quad (\mathbf{f} \ \mathbf{g})U = (\mathbf{f} \ \mathbf{m}).$$

Since the first column on the right is $\mathbf{f}$, we know the first column in U is $\mathbf{a}$. If the second column in U is (b, c) , then $c = \pm 1$ since $\det U = \pm 1$.

Equating the second columns in (*), we find $b\mathbf{f} = -c\mathbf{g} + \mathbf{m}$. Now take $b = -ac$: $-ac\mathbf{f} = -c\mathbf{g} + \mathbf{m}$, or $a\mathbf{f} = \mathbf{g} - c\mathbf{m}$. Since $c = \pm 1$, this completes the exercise.

46-4. First, consider the special case when one vector (P_1, P_2) is on the positive y-axis (so $P_1 = 0$), and the other vector is (Q_1, Q_2) . The length of the base is P_2 , and the altitude is $|Q_1|$, so the area is

$$|P_2 Q_1| = \left| \det \begin{pmatrix} P_1 & Q_1 \\ P_2 & Q_2 \end{pmatrix} \right|.$$

Now suppose $\mathbf{f} = (p_1, p_2)$ and $\mathbf{g} = (q_1, q_2)$ are in general position. If we make a rotation of axes through an appropriate angle θ so the $\mathbf{f}$ vector is on the positive y-axis of the new coordinate system, the new coordinates (P_1, P_2) and (Q_1, Q_2) are given by

$$\begin{pmatrix} P_1 & Q_1 \\ P_2 & Q_2 \end{pmatrix} = T(\mathbf{f} \ \mathbf{g}), \text{ where } T = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \text{ The area is}$$

$\left| \det \begin{pmatrix} P_1 & Q_1 \\ P_2 & Q_2 \end{pmatrix} \right| = |\det(T(\mathbf{f} \ \mathbf{g}))| = |\det T| |\det(\mathbf{f} \ \mathbf{g})|$. But since $\det T = 1$, we are through.

46-5. Since $m, n \in \Lambda$, there are integers a, b, c, d such that $m = af + bg$ and $n = cf + dg$. Thus, $(m, n) = (f, g) \begin{pmatrix} a & c \\ b & d \end{pmatrix}$. Taking determinants we have

$$|\det(m, n)| = |\det(f, g)| \left| \det \begin{pmatrix} a & c \\ b & d \end{pmatrix} \right|. \text{ But since } |\det(m, n)| = |\det(f, g)|, \text{ it follows that}$$

$$\left| \det \begin{pmatrix} a & c \\ b & d \end{pmatrix} \right| = \pm 1. \text{ With } U = \begin{pmatrix} a & c \\ b & d \end{pmatrix}, \text{ by Theorem 46.1 we have } m, n = f, g.$$

46-6. Suppose there is a point $z \in \Lambda(f, g)$ which is inside the triangle o, m, n . Then $z = xm + yn$ for some real x, y such that $0 \leq x \leq 1, 0 \leq y \leq 1$, and $(x, y) \notin \{o, a, b\}$. Therefore, $z \notin \Lambda(m, n)$, so $m, n \neq f, g$.

Now assume $m, n \neq f, g$. Since $m, n \in \Lambda(f, g)$, we have $\Lambda(m, n) \subset \Lambda(f, g)$. But because $\Lambda(m, n) \neq \Lambda(f, g)$, there is a point $z \in \Lambda(f, g)$ such that $z \notin \Lambda(m, n)$. Say $z = xm + yn$, where at least one of x, y is not an integer. Now consider the point $s = (x - [x])m + (y - [y])n$. Both $x - [x]$ and $y - [y]$ are non-negative, both are < 1 , and not both are 0. Therefore, s is in triangle o, m, n , and s is not a vertex of the triangle. But $s = z - ([x]m + [y]n) \in \Lambda(f, g)$.

46-7. In Figure 6, consider a line parallel to L_0 and halfway between L_0 and L_1 . Clearly such a line contains no points of the lattice.

A line through o containing the vector $f + (\sqrt{2})g$ has equation $tf + (t\sqrt{2})g$ for t real, $-\infty < t < \infty$. Such a line cannot contain any point of the lattice except o , because if it did, then for some real t we would have $tf + (t\sqrt{2})g = af + bg$ for integers a and b , which is impossible.

Finally, suppose L is a line containing two points $z = af + bg$ and $s = cf + dg$. The line through z and s has equation $z + t(s - z)$ for real t , $-\infty < t < \infty$. Whenever t is an integer, $z + t(s - z) = ((1 + t)a - tc)f + ((1 + t)b - td)g$ which is in Λ .

Section 47

47-1. Suppose $\mathcal{U} = \bigcup_{\alpha \in \mathfrak{A}} \Omega_\alpha$, Ω_α open, where $\mathfrak{A}$ is an index set. Let f be any point in $\mathcal{U}$.

Then for some $\alpha \in \mathfrak{A}$, $f \in \Omega_\alpha$. Since Ω_α is open, f is an interior point of Ω_α . Therefore, every point of $\mathcal{U}$ is an interior point, so $\mathcal{U}$ is open.

Now let $\mathcal{F} = \bigcup_{n=1}^N \mathcal{F}_n$, $\mathcal{F}_n$ closed. Let f be in the complement of $\mathcal{F}$. Then $f \notin \mathcal{F}_n$

for every n , $n = 1, \dots, N$. For each n , there is a circle C_n with center at f , radius r_n , such that $C_n \cap \mathcal{F}_n = \emptyset$. Let $r = \min r_n$. Then the circle C with center at f and radius r contains no points of $\mathcal{F}_1$, no points of $\mathcal{F}_2$, ..., no points of $\mathcal{F}_N$. That is, C lies entirely inside the complement of $\mathcal{F}$. Thus, $\mathcal{F}$ is closed.

Taking complements tells us that the intersection of any number of closed sets is closed, and the intersection of a finite number of open sets is open.

47-3. $f(\mathbf{p}) = \max(|p_1|, |p_2|) \geq 0$, and $= 0$ if and only if $p_1 = p_2 = 0$. Also, $f(\mathbf{p}) = \max(|p_1|, |p_2|) = |\mathbf{p}| \max(|p_1|, |p_2|) = |\mathbf{p}| f(\mathbf{p})$. For the triangle inequality, assume first that $|p_1 + q_1| \leq |p_2 + q_2|$. Then $f(\mathbf{p} + \mathbf{q}) = |p_2 + q_2| \leq |p_2| + |q_2| \leq \max(|p_1|, |p_2|) + \max(|q_1|, |q_2|) = f(\mathbf{p}) + f(\mathbf{q})$. The proof is similar if $|p_2 + q_2| < |p_1 + q_1|$.

The set $f(\mathbf{p}) \leq 1$ is the closed square with vertices $(1, 1)$, $(-1, 1)$, $(-1, -1)$, and $(1, -1)$.

47-5. Define $f(\mathbf{o}) = 0$. If $\mathbf{p} \neq \mathbf{o}$, there is a unique scalar c , c is a positive real number, such that $c\mathbf{p}$ is on the boundary of $\mathcal{S}$. (The vector $c\mathbf{p}$ is a vector with the same direction as $\mathbf{p}$, but the length is changed by a factor c . In general, $c\mathbf{p}$ is not a lattice point in $\Lambda(\mathbf{a}, \mathbf{b})$.) Define $f(\mathbf{p}) = 1/c$. Notice that if $\mathbf{p} \in \mathcal{S}$, then $c \geq 1$, so $f(\mathbf{p}) \leq 1$; if $\mathbf{p} \notin \mathcal{S}$, $c < 1$, so $f(\mathbf{p}) > 1$.

Specifically, if $\mathbf{p} = (e, d)$ (or its extension) intersects the left or right side of the parallelogram, then $f(\mathbf{p}) = |6e - d|/16$; if $\mathbf{p}$ or its extension intersects the top or bottom of the parallelogram, then $f(\mathbf{p}) = |2e + 5d|/16$.

Section 48

48-1. With the metric $f(\mathbf{p}) = |\mathbf{p}|$, the set $\mathcal{S}$ is a circle with radius 1 and center $\mathbf{o}$. If $\mathbf{a}, \mathbf{b}$ are any vectors, then $2\Lambda(\mathbf{a}, \mathbf{b}) = \{2j\mathbf{a} + 2k\mathbf{b} : j, k \in \mathbb{Z}\}$. The collection of sets $\mathcal{S} \oplus \mathbf{p}$ is the collection of circles of radius 1 with centers in 2Λ . Thus, if $\min(|\mathbf{a}|, |\mathbf{b}|) > 1$, no two circles intersect; if the minimum $= 1$, the circles intersect at some points in Λ , but the points of intersection are on the boundary of the circles, and there are no interior points in the intersections; if the minimum length < 1 , then odd multiples of the vector $(\mathbf{a}$ or $\mathbf{b})$ with minimum length are interior points in the intersections. Therefore, all of (a), (b), and (c) are $\mathcal{S}$ -admissible.

48-3. (a) $\mathcal{S} \cap \Lambda = \{x\mathbf{a} + y\mathbf{b} : x \in \{0, \pm 1/2, \pm 1\}, y \in \{0, \pm 1\}\}$.

(b) $\mathcal{S} \cap \Lambda = \mathbf{o}$.

(c) $\mathcal{S} \cap \Lambda = \mathbf{o}$.

48-5. If $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$, notice that $F(\mathbf{p}) = u^2 + v^2$ and $F(\mathbf{q}) = w^2 + z^2$, with $u = \sqrt{\alpha}p_1 + \frac{\beta p_2}{\sqrt{\alpha}}$, $v = \frac{\sqrt{\delta}p_2}{\sqrt{\alpha}}$, $w = \sqrt{\alpha}q_1 + \frac{\beta q_2}{\sqrt{\alpha}}$, and $z = \frac{\sqrt{\delta}q_2}{\sqrt{\alpha}}$. Then $F(\mathbf{p} + \mathbf{q}) = (u + w)^2 + (v + z)^2$. Now the details of showing that $f(\mathbf{p}) = (F(\mathbf{p}))^{1/2}$ is a csdf are precisely the same as those showing that $g(\mathbf{p}) = |\mathbf{p}|$ is a csdf.

Section 49

49-1. For Corollary 33.1a, suppose $f = (b, a)$, $g = (d, c)$, and $a/b < c/d$ are successive. We have $\det(fg) = bc - ad = \pm 1$. But $a/b < c/d$ says $bc - ad > 0$, so $bc - ad = 1$.

For Corollary 33.1b, we have $a/b < u/v$, $u/v < c/d$, and $|\det(fm)| = |\det(mg)|$. Therefore, $|bu - av| = |cv - du|$. But $a/b < u/v$ says $0 < bu - av$; similarly, $u/v < c/d$ says $cv - du > 0$. Thus, $bu - av = cv - du$, and $u/v = (a + c)/(b + d)$.

49-3. If $x_1 < x_2$, then $T(x_1) > T(x_2)$, so the transformation is monotonically decreasing.

The fixed point is $\frac{\alpha - \delta + \sqrt{(\alpha + \delta)^2 + 4}}{2\gamma}$. We again have $|T(x_1) - x_0| < |x_1 - x_0|/2$,

where x_0 is the fixed point; therefore, all sequences $\{T^k(x_1)\}$ converge to the fixed point.

BIBLIOGRAPHY

1. Harold Davenport, *The Higher Arithmetic*. Harper, New York, 1960.
2. Emil Grosswald, *Topics from the Theory of Numbers*. Macmillan, New York, 1966.
3. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 4th ed. Oxford University Press, London, 1959.
4. William J. LeVeque, *Topics in Number Theory*, Vol. I. Addison-Wesley, Reading, Mass., 1956.
5. Kurt Mahler, *Lecture Notes on Geometric Number Theory*. Notre Dame University, 1962.
6. Trygve Nagell, *Introduction to Number Theory*, 2nd ed. Chelsea, New York, 1964.
7. Ivan Niven and Herbert S. Zuckerman, *An Introduction to the Theory of Numbers*, 2nd ed. Wiley, New York, 1966.
8. Hans Rademacher, *Lectures on Elementary Number Theory*. Blaisdell, Waltham, Mass., 1964.
9. Harold N. Shapiro, *Lecture Notes on the Theory of Numbers*. New York University, 1951–1952.
10. James E. Shockley, *Introduction to Number Theory*. Holt, Rinehart and Winston, New York, 1967.

INDEX

- Admissible lattice, 170
- Algorithm
 - continued fraction, 100, 105, 179
 - division ($\mathcal{G}$), 91
 - division ($\mathcal{J}$), 125
 - division ($\mathcal{Z}$), 2, 4, 7, 8, 31, 34, 41, 57
 - Euclidean, 8, 36, 92, 126
- Amicable numbers, 26
- Arithmetic function, 13
- Associates, 91, 124
- Associative law, 14
- Asymptotic equivalence, 64
- Belongs to, 47
- Bertrand's conjecture, 156
- Big- O , 64
- Bounded set, 165
- Canonical form, 10
- Characteristic function, 22
- Chinese Remainder Theorem, 37, 61, 62
- Closed set, 165
- Commutative law, 14
- Completely multiplicative, 28, 29
- Complete residue system, 34, 41, 87
- Composite, 10
- Congruence (mod m), 31
- Conjugate, 91
- Continued fraction, 98
 - algorithm, 100, 105, 179
 - convergent to, 101, 117, 118, 175
 - finite, 98
 - infinite, 103, 179
 - irrational value, 104, 179
 - partial quotient of, 99
 - periodic, 106, 175, 179
 - rational value, 100
 - simple, 99
- Convergent (*See* Continued fraction)
- Convex body, 165
- Convex set, 165
- Convex symmetric distance function, 166
- Convolution product, 13
- CRS (mod m), 34, 41, 87
- csdf, 166
- Dependent, linearly, 161
- Diophantine equation
 - definition, 35
 - $ax + by = n$, 36
 - $x^2 + y^2 = n$, 87, 88, 89
 - $x^2 + y^2 = z^2$, 121
 - $x^2 + 3y^2 = n$, 124
 - $x^2 - Dy^2 = 1$, 98, 110
 - $x^2 - Dy^2 = -1$, 114
 - $x^3 + y^3 = z^3$, 128
 - $x^4 + y^4 = z^2$, 122
 - $x^4 + y^4 = z^4$, 123
 - $u^2 + v^2 + x^2 + y^2 = n$, 84, 86
- Dirichlet convolution, 13
- Dirichlet series, 80
- Divides, 1, 91
- Divisibility, 1
- Division algorithm (*See* Algorithm)
- Dominant factor, 111
- Equivalence, asymptotic, 64
- Equivalence class, 32
- Equivalence relation, 31, 32, 34
- Equivalent point-pairs, 162, 163
- Euclidean algorithm, 8, 36, 92, 126
- Euler-McLaurin sum formula, 71
- Euler's constant, 66, 72, 74–76
- Euler's criterion, 53, 87
- Euler's ϕ -function (*See* Functions)

- Euler's Theorem, 39, 40, 46, 49
- Euler's totient (*See* Functions)
- Exponent, 47
- Exponent, universal, 49

- Factor, dominant, 111
- Factor, fundamental, 111
- Factor, inferior, 111
- Factorization, 10, 12, 91, 93, 126
- Farey sequence, 107, 114, 115, 174, 175
- Fermat's Last Theorem, 120
- Fermat's Theorem, 40, 47, 62
- Fibonacci sequence, 106
- Field, 34
- Finite continued fraction, 98
- Fixed point, 177
- Fraction, continued (*See* Continued fraction)
- Functions
 - arithmetic, 13
 - characteristic, 22
 - completely multiplicative, 28, 29
 - csdf, 166
 - Euler's, 27, 39–41, 43, 48–52, 78, 82, 107
 - inverse, 16, 21, 23
 - iota, 18–20, 23, 45, 77, 78, 81, 82, 130
 - Jordan's, 29
 - largest integer, 15
 - Möbius, 21, 22, 44, 78, 79, 81, 83, 130, 132, 133
 - multiplicative, 18, 20, 22, 24, 28, 29, 43
 - natural product of, 65, 130
 - product of, 13
 - Ramanujan's, 42, 82
 - Riemann zeta, 81, 82
 - sigma, 23, 24, 73, 75, 77, 82
 - square bracket, 15
 - sum of, 65, 130
 - summatory, 63
 - totient, 28, 29
- Fundamental factor, 111
- Fundamental solution, 111

- Gaussian integer, 90
- Gauss Lemma, 54, 57, 58
- gcd, 4, 6, 7, 11, 92, 94

- Greatest common divisor, 4, 6, 7, 11, 92, 94
- Group, 17

- Hurwitz' Theorem, 115

- Idempotent, 16
- Identity, 14
- Iota functions (*See* Functions)
- Imprimitive, 87
- Independent, linearly, 161
- Inferior factor, 111
- Infinite continued fraction, 103, 179
- Integers
 - Gaussian, 90
 - Jacobian, 123
 - modulo m , 33
 - rational, 1
- Integral lattice, 161
- Integral matrix, 163
- Integral module, 7
- Interior point, 165
- Inverse function, 16, 21, 23

- Jacobian integers, 123
- Jordan totient, 29

- Lagrange's Theorem, 45, 48
- Largest integer function, 15
- Lattice, admissible, 170
- Lattice, integral, 161
- Lattice point, 55, 56, 73, 79, 161
- lcm, 5, 7, 11
- Least common multiple, 5, 7, 11
- Legendre symbol, 53, 87, 94, 124
- Linear combination, 3, 161
- Linearly dependent, 161
- Linearly independent, 161
- Little- o , 64
- Logarithmic integral, 157

- Matrix, integral, 163
- Mediant, 107
- Mersenne prime, 25

- Minkowski's Theorem, 171, 172
- Möbius function (*See* Functions)
- Möbius inversion, 22, 25, 26, 28, 45, 131
- Module, integral, 7
- Modulus, 31, 92
- Multiple, 1
- Multiplicative, 18, 20, 22, 24, 28, 29, 43
- Multiplicative, completely, 28, 29

- Natural product, 65, 130
- Norm, 90, 123

- Open set, 165

- Partial order, 3
- Partial quotient, 99
- Partition, 28
- Pell equation, 110
 - dominant factor, 111
 - fundamental factor, 111
 - fundamental solution, 111
 - inferior factor, 111
 - positive solution, 110
- Perfect number, 25
- Periodic continued fraction, 106, 175, 179
- Positive solution, 110
- Prime, 9, 12, 91, 93, 124
- Prime Number Theorem, 129, 153–155
- Prime, relatively, 5
- Primitive representation, 87
- Primitive root, 48, 50–52
- Primitive triple, 120
- Product, convolution, 13
- Product, natural, 65, 130
- Pythagorean triple, 120

- Quadratic irrational, 175, 179
- Quadratic non-residue, 53
- Quadratic reciprocity, 57
- Quadratic residue, 53
- Quotient, partial, 99

- Ramanujan's function, 42, 82
- Ramify, 94, 126
- Reciprocity law, 57
- Reduced residue system, 39, 42

- Reflexive relation, 32
- Relation
 - equivalence, 31, 32, 34
 - reflexive, 32
 - symmetric, 32
 - transitive, 32
- Relatively prime, 5
- Residue, 31
- Residue class, 33
- Riemann zeta function, 81, 82
- Ring, 34
- RRS (mod m), 39, 42

- Selberg's formula, 139
- Semigroup, 14
- Set
 - bounded, 165
 - closed, 165
 - convex, 165
 - open, 165
 - symmetric, 165
- Sigma functions (*See* Functions)
- Simple continued fraction, 99
- Solution point, 39
- Split, 94, 126
- Square bracket function, 15
- Squarefree number, 22, 82
- Successive, 107
- Successor, 107
- Sum formula, 71
- Sum of functions, 65, 130
- Summation by parts, 134
- Summatory function, 63
- Symmetric relation, 39
- Symmetric set, 165

- Totient function, 28, 29
- Transitive relation, 32

- Unique factorization, 10, 91, 93, 126
- Unit, 17, 90, 123
- Universal exponent, 49

- Wilson's Theorem, 46, 47

- Zeta function, 81, 82

A CATALOG OF SELECTED DOVER BOOKS IN ALL FIELDS OF INTEREST

CONCERNING THE SPIRITUAL IN ART, Wassily Kandinsky. Pioneering work by father of abstract art. Thoughts on color theory, nature of art. Analysis of earlier masters. 12 illustrations. 80pp. of text. 5% x 8%. 23411-8 Pa. \$4.95

ANIMALS: 1,419 Copyright-Free Illustrations of Mammals, Birds, Fish, Insects, etc., Jim Harter (ed.). Clear wood engravings present, in extremely lifelike poses, over 1,000 species of animals. One of the most extensive pictorial sourcebooks of its kind. Captions. Index. 284pp. 9 x 12. 23766-4 Pa. \$14.95

CELTIC ART: The Methods of Construction, George Bain. Simple geometric techniques for making Celtic interlacements, spirals, Kells-type initials, animals, humans, etc. Over 500 illustrations. 160pp. 9 x 12. (Available in U.S. only.) 22923-8 Pa. \$9.95

AN ATLAS OF ANATOMY FOR ARTISTS, Fritz Schider. Most thorough reference work on art anatomy in the world. Hundreds of illustrations, including selections from works by Vesalius, Leonardo, Goya, Ingres, Michelangelo, others. 593 illustrations. 192pp. 7% x 10%. 20241-0 Pa. \$9.95

CELTIC HAND STROKE-BY-STROKE (Irish Half-Uncial from "The Book of Kells"): An Arthur Baker Calligraphy Manual, Arthur Baker. Complete guide to creating each letter of the alphabet in distinctive Celtic manner. Covers hand position, strokes, pens, inks, paper, more. Illustrated. 48pp. 8% x 11. 24336-2 Pa. \$3.95

EASY ORIGAMI, John Montroll. Charming collection of 32 projects (hat, cup, pelican, piano, swan, many more) specially designed for the novice origami hobbyist. Clearly illustrated easy-to-follow instructions insure that even beginning paper-crafters will achieve successful results. 48pp. 8% x 11. 27298-2 Pa. \$3.50

THE COMPLETE BOOK OF BIRDHOUSE CONSTRUCTION FOR WOODWORKERS, Scott D. Campbell. Detailed instructions, illustrations, tables. Also data on bird habitat and instinct patterns. Bibliography. 3 tables. 63 illustrations in 15 figures. 48pp. 5% x 8%. 24407-5 Pa. \$2.50

BLOOMINGDALE'S ILLUSTRATED 1886 CATALOG: Fashions, Dry Goods and Housewares, Bloomingdale Brothers. Famed merchants' extremely rare catalog depicting about 1,700 products: clothing, housewares, firearms, dry goods, jewelry, more. Invaluable for dating, identifying vintage items. Also, copyright-free graphics for artists, designers. Co-published with Henry Ford Museum & Greenfield Village. 160pp. 8% x 11. 25780-0 Pa. \$10.95

HISTORIC COSTUME IN PICTURES, Braun & Schneider. Over 1,450 costumed figures in clearly detailed engravings—from dawn of civilization to end of 19th century. Captions. Many folk costumes. 256pp. 8% x 11%. 23150-X Pa. \$12.95

CATALOG OF DOVER BOOKS

PIANO TUNING, J. Cree Fischer. Clearest, best book for beginner, amateur. Simple repairs, raising dropped notes, tuning by easy method of flattened fifths. No previous skills needed. 4 illustrations. 201pp. 5% x 8%. 23267-0 Pa. \$6.95

HINTS TO SINGERS, Lillian Nordica. Selecting the right teacher, developing confidence, overcoming stage fright, and many other important skills receive thoughtful discussion in this indispensable guide, written by a world-famous diva of four decades' experience. 96pp. 5³/₈ x 8¹/₂. 40094-8 Pa. \$4.95

THE COMPLETE NONSENSE OF EDWARD LEAR, Edward Lear. All nonsense limericks, zany alphabets, Owl and Pussycat, songs, nonsense botany, etc., illustrated by Lear. Total of 320pp. 5% x 8%. (AVAILABLE IN U.S. ONLY.) 20167-8 Pa. \$7.95

VICTORIAN PARLOUR POETRY: An Annotated Anthology, Michael R. Turner. 117 gems by Longfellow, Tennyson, Browning, many lesser-known poets. "The Village Blacksmith," "Curfew Must Not Ring Tonight," "Only a Baby Small," dozens more, often difficult to find elsewhere. Index of poets, titles, first lines. xxiii + 325pp. 5% x 8%. 27044-0 Pa. \$8.95

DUBLINERS, James Joyce. Fifteen stories offer vivid, tightly focused observations of the lives of Dublin's poorer classes. At least one, "The Dead," is considered a masterpiece. Reprinted complete and unabridged from standard edition. 160pp. 5% x 8%. 26870-5 Pa. \$1.00

GREAT WEIRD TALES: 14 Stories by Lovecraft, Blackwood, Machen and Others, S. T. Joshi (ed.). 14 spellbinding tales, including "The Sin Eater," by Fiona McLeod, "The Eye Above the Mantel," by Frank Belknap Long, as well as renowned works by R. H. Barlow, Lord Dunsany, Arthur Machen, W. C. Morrow and eight other masters of the genre. 256pp. 5% x 8%. (Available in U.S. only.) 40436-6 Pa. \$8.95

THE BOOK OF THE SACRED MAGIC OF ABRAHELIN THE MAGE, translated by S. MacGregor Mathers. Medieval manuscript of ceremonial magic. Basic document in Aleister Crowley, Golden Dawn groups. 268pp. 5% x 8%. 23211-5 Pa. \$9.95

NEW RUSSIAN-ENGLISH AND ENGLISH-RUSSIAN DICTIONARY, M. A. O'Brien. This is a remarkably handy Russian dictionary, containing a surprising amount of information, including over 70,000 entries. 366pp. 4¹/₂ x 6%. 20208-9 Pa. \$10.95

HISTORIC HOMES OF THE AMERICAN PRESIDENTS, Second, Revised Edition, Irvin Haas. A traveler's guide to American Presidential homes, most open to the public, depicting and describing homes occupied by every American President from George Washington to George Bush. With visiting hours, admission charges, travel routes. 175 photographs. Index. 160pp. 8% x 11. 26751-2 Pa. \$11.95

NEW YORK IN THE FORTIES, Andreas Feininger. 162 brilliant photographs by the well-known photographer, formerly with *Life* magazine. Commuters, shoppers, Times Square at night, much else from city at its peak. Captions by John von Hartz. 181pp. 9% x 10%. 23585-8 Pa. \$13.95

INDIAN SIGN LANGUAGE, William Tomkins. Over 525 signs developed by Sioux and other tribes. Written instructions and diagrams. Also 290 pictographs. 111pp. 6% x 9%. 22029-X Pa. \$3.95

CATALOG OF DOVER BOOKS

THE INFLUENCE OF SEA POWER UPON HISTORY, 1660-1783, A. T. Mahan. Influential classic of naval history and tactics still used as text in war colleges. First paperback edition. 4 maps. 24 battle plans. 640pp. 5% x 8%. 25509-3 Pa. \$14.95

THE STORY OF THE TITANIC AS TOLD BY ITS SURVIVORS, Jack Winocour (ed.). What it was really like. Panic, despair, shocking inefficiency, and a little heroism. More thrilling than any fictional account. 26 illustrations. 320pp. 5% x 8%. 20610-6 Pa. \$8.95

FAIRY AND FOLK TALES OF THE IRISH PEASANTRY, William Butler Yeats (ed.). Treasury of 64 tales from the twilight world of Celtic myth and legend: "The Soul Cages," "The Kildare Pooka," "King O'Toole and his Goose," many more. Introduction and Notes by W. B. Yeats. 352pp. 5% x 8%. 26941-8 Pa. \$8.95

BUDDHIST MAHAYANA TEXTS, E. B. Cowell and others (eds.). Superb, accurate translations of basic documents in Mahayana Buddhism, highly important in history of religions. The Buddha-karita of Asvaghosha, Larger Sukhavativyuha, more. 448pp. 5% x 8%. 25552-2 Pa. \$12.95

ONE TWO THREE . . . INFINITY: Facts and Speculations of Science, George Gamow. Great physicist's fascinating, readable overview of contemporary science: number theory, relativity, fourth dimension, entropy, genes, atomic structure, much more. 128 illustrations. Index. 352pp. 5% x 8%. 25664-2 Pa. \$9.95

EXPERIMENTATION AND MEASUREMENT, W. J. Youden. Introductory manual explains laws of measurement in simple terms and offers tips for achieving accuracy and minimizing errors. Mathematics of measurement, use of instruments, experimenting with machines. 1994 edition. Foreword. Preface. Introduction. Epilogue. Selected Readings. Glossary. Index. Tables and figures. 128pp. 5³/₈ x 8¹/₂. 40451-X Pa. \$6.95

DALÍ ON MODERN ART: The Cuckolds of Antiquated Modern Art, Salvador Dalí. Influential painter skewers modern art and its practitioners. Outrageous evaluations of Picasso, Cézanne, Turner, more. 15 renderings of paintings discussed. 44 calligraphic decorations by Dalí. 96pp. 5% x 8%. (Available in U.S. only.) 29220-7 Pa. \$5.95

ANTIQUATE PLAYING CARDS: A Pictorial History, Henry René D'Allemagne. Over 900 elaborate, decorative images from rare playing cards (14th-20th centuries): Bacchus, death, dancing dogs, hunting scenes, royal coats of arms, players cheating, much more. 96pp. 9% x 12%. 29265-7 Pa. \$12.95

MAKING FURNITURE MASTERPIECES: 30 Projects with Measured Drawings, Franklin H. Gottshall. Step-by-step instructions, illustrations for constructing handsome, useful pieces, among them a Sheraton desk, Chippendale chair, Spanish desk, Queen Anne table and a William and Mary dressing mirror. 224pp. 8% x 11%. 29338-6 Pa. \$13.95

THE FOSSIL BOOK: A Record of Prehistoric Life, Patricia V. Rich et al. Profusely illustrated definitive guide covers everything from single-celled organisms and dinosaurs to birds and mammals and the interplay between climate and man. Over 1,500 illustrations. 760pp. 7% x 10%. 29371-8 Pa. \$29.95

Prices subject to change without notice.

Available at your book dealer or write for free catalog to Dept. GI, Dover Publications, Inc., 31 East 2nd St., Mineola, N.Y. 11501. Dover publishes more than 500 books each year on science, elementary and advanced mathematics, biology, music, art, literary history, social sciences and other areas.